SRAO: COORDINATING AUXILIARY GUIDANCE AND FEEDBACK FOR EFFICIENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning systems often combine inexpensive guidance with feedback that is costly to obtain. Their efficiency depends on coordinating how guidance shapes an update with how feedback corrects it. We introduce Shared Residual Audited Optimization (SRAO), which uses corrective feedback both to improve the current update and to inform the use of guidance. For selective gradient acquisition, one reference corrects the estimate and evaluates all available guidance-acquisition rules. The key insight is that the estimation gain from combining guidance and acquisition choices can pay for learning how to combine them. We establish this connection through a sharp variance bound and a cumulative learning guarantee. Scalar and gradient studies explain the mechanism, including about 40% lower estimation error at matched reference cost. On-policy distillation retains similar observed mean accuracy with 35.5% fewer reference-gradient batches. Agentic question answering extends the coordination principle to reflection guidance and outcome feedback, improving mean task quality. Together, these results connect feedback efficiency to the way guidance enters learning updates.

1 INTRODUCTION

Learning algorithms often have useful guidance before they can obtain a fresh reference signal. Cached computations, model predictions and auxiliary objectives suggest updates at different costs and levels of reliability. Online learning exploits predictable information (Rakhlin & Sridharan, 2013), multifidelity estimation combines cheap approximations with expensive observations (Peherstorfer et al., 2016), and auxiliary learning adapts additional objectives to a primary task (Du et al., 2018). Language-model training makes this problem especially visible. On-policy distillation acquires teacher supervision for student-generated responses (Agarwal et al., 2024), while process supervision and outcome rewards provide feedback at different levels of a reasoning trajectory (Lightman et al., 2024). Across these settings, the learner must decide how guidance should influence an update and how feedback should correct it.

The common difficulty is that these decisions are coupled. A cached gradient may become stale as the model changes, and an auxiliary objective may help some updates while opposing others. Fresh feedback reveals the discrepancy, but can require substantial inference or gradient computation. Selective observation is therefore attractive, as in active preference learning (Muldrew et al., 2024). Yet acquiring less feedback also makes the reliability of guidance harder to assess. *How can each corrective observation improve both the current update and the learner's use of guidance?*

We introduce Shared Residual Audited Optimization (SRAO) around one principle: **use corrective feedback to determine how guidance enters the update**. For selective gradient acquisition, each rule pairs a prediction of a common reference with a probability of acquiring it. SRAO combines these rules conditional on the acquisition outcome. Rules that query more often contribute less to the prediction used without a query. The resulting estimate preserves the reference mean and expected acquisition cost while removing the error caused by randomly choosing a rule. One acquired reference then evaluates every rule through its residual, supplying shared evidence for learning the guidance and acquisition weights. Figure 1 shows the common guidance–feedback design.

The technical challenge is to retain the estimation gain while learning from selective feedback. Active inference already adapts predictions and acquisition (Zrnic & Candes, 2024), and online variance reduction learns sampling mixtures (Borsos et al., 2019). SRAO's central contribution is *a sharp*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

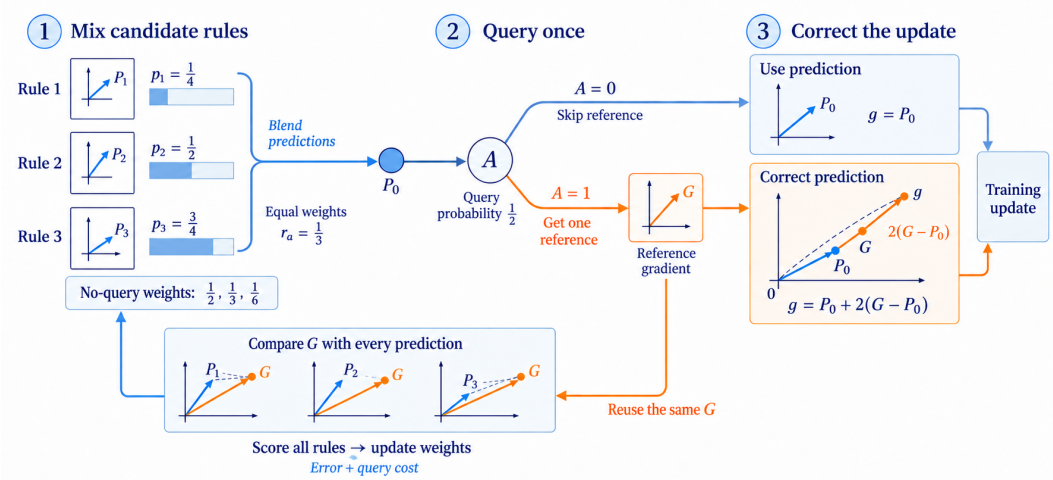


Figure 1: **SRAO uses corrective feedback to regulate guidance.** Guidance supplies a candidate direction, and feedback determines how it enters the training update. In on-policy distillation, acquired reference gradients correct cached-gradient guidance. In agentic question answering, outcome feedback regulates reflection- and backtracking-guided optimizer displacements. The two applications instantiate the same principle with their respective correction operators.

relation between the error removed from the executed estimate and the variance cost of learning the rule weights. This relation lets the estimator's own gain pay for adaptation. Section 2 positions the construction, with broader connections in Appendix C.

For a fixed finite library, bounded residuals and a positive query floor, the resulting expected cumulative excess in estimation error plus reference cost is bounded independently of the training horizon. The comparator is the best fixed rule identity in expectation on the learner's history, where each rule may respond to the current context. An exact error decomposition then separates selective acquisition from the reference's own bias and noise, yielding a consequence for smooth gradient optimization. The analysis concerns a shared reference and available predictions, independently of the architecture or task.

The experiments connect this statistical result to practical learning. Scalar studies show how the estimation gain can offset the cost of learning. Controlled language-model gradients exhibit about 40% lower MSE at matched reference cost. Separate training histories identify the gain from combining distinct predictions. In on-policy distillation with the OPSD objective (Zhao et al., 2026), SRAO retains similar observed mean accuracy with 35.5% fewer reference-gradient batches. Agentic question answering uses outcome feedback to regulate reflection-guided optimizer displacements, extending the principle beyond selective acquisition. The full system improves mean answer quality, and removing its coordination stage lowers the seven-task mean by 4.94 points at 3B. This application is evaluated empirically, outside the selective-gradient guarantee.

2 Related Work

SRAO connects prediction-based estimation and selective-feedback learning through a sharp relation between estimator gain and the cost of learning guidance–acquisition weights.

Augmented estimators correct predictions using selectively observed residuals (Robins et al., 1994). Active statistical inference adapts prediction and acquisition jointly (Zrnic & Candes, 2024). SRAO uses a finite library whose rules share one reference, then conditions their aggregate on the acquisition outcome. This classical variance reduction supplies the gain analyzed by our sharp inequality.

Online variance reduction learns sampling mixtures along an optimization trajectory (Borsos et al., 2019). AdaHedge controls the cost of learning through observed loss variance (de Rooij et al., 2014). Negative variance terms can compensate such costs (van der Hoeven et al., 2022). SRAO connects these ideas through an actual reduction in the executed estimator's error, covering both

query outcomes and heterogeneous predictions. In the full-feedback application, auxiliary-gradient coordination (Du et al., 2018) and conflict projection (Yu et al., 2020) provide related ways to regulate guidance. SRAO's agent operator acts on optimizer displacements after look-ahead and outcome rescoring. Appendix C gives the broader comparison.

3 SRAO: Coordinating Guidance and Feedback

SRAO uses one corrective observation both to improve an update and to assess the guidance available to the learner. For selective gradient acquisition, conditional averaging gives a better estimate at the same expected reference cost. The acquired residual then evaluates every rule, allowing one learning update to coordinate guidance and acquisition.

3.1 Guidance and correction

At round t, the common evaluation point y_t is fixed by the training protocol. The available history $\mathcal{H}_t$ may include generated responses, rewards and features. Each of $K \geq 2$ feedback rules supplies $(P_{t,a}, p_{t,a})$, where $P_{t,a}$ predicts the same reference gradient G_t at y_t and $p_{t,a} \in [\varepsilon, 1]$, $\varepsilon > 0$. The weights satisfy $r_{t,a} \geq 0$ and $\sum_a r_{t,a} = 1$. Rule identity is fixed throughout the run, while its prediction and probability may respond to the history and current context. Feedback cost $c_t \geq 0$ measures the complete reference computation in a declared unit. The conversion $\kappa_t = \lambda c_t$ for $\lambda \geq 0$ expresses this cost in the units used to trade it against squared estimation error.

The acquisition protocol matters because feedback must test choices made before seeing the reference. Predictions, probabilities, costs and weights are all fixed before the query. Conditional on $\mathcal{H}_t$, the fresh query coin is independent of the potential reference G_t. Write $\mathbb{E}_t = \mathbb{E}[\cdot \mid \mathcal{H}_t]$. Vector mean identities assume integrability of the reference and predictions. The error comparisons use the separate residual-moment conditions stated in Section 4.1.

One way to use the weights is to sample a rule $I_t \sim r_t$, query with probability p_{t,I_t}, and return $P_{t,I_t} + A_t(G_t - P_{t,I_t})/p_{t,I_t}$. Inverse-probability weighting corrects the sampled rule's residual (Horvitz & Thompson, 1952). SRAO removes the extra randomness from choosing the rule while retaining the query decision:

$$\bar{p}_t = \sum_a r_{t,a} p_{t,a}, \qquad P_{0,t} = \frac{\sum_a r_{t,a}(1 - p_{t,a}) P_{t,a}}{1 - \bar{p}_t}, \qquad g_t = P_{0,t} + \frac{A_t}{\bar{p}_t}(G_t - P_{0,t}). \quad (1)$$

Here $A_t \sim \mathrm{Bernoulli}(\bar{p}_t)$. If $\bar{p}_t = 1$, set $g_t = G_t$ without defining $P_{0,t}$. Otherwise, $P_{0,t}$ averages the predictions conditional on *no query*. A rule that queries more often receives less weight in the estimate used without feedback. The result generally differs from the ordinary average $\sum_a r_{t,a} P_{t,a}$. The reason is informational: *not querying also changes which rule was likely to have been selected*. Two equally weighted rules with query probabilities $1/4$ and $3/4$ receive weights $3/4$ and $1/4$ after a no-query decision. Figure 2(a) shows how this changes the prediction and its error.

Equation (1) is the conditional expectation of the sampled-rule estimate given the query outcome and any acquired reference. This Rao–Blackwellization preserves the reference mean, $\mathbb{E}_t g_t = \mathbb{E}_t G_t$, and expected feedback cost with no greater estimation error. The state-update rule acts on g_t, producing the history on which the rules are compared.

Fixing all weight on one rule recovers the residual-correction update $g_t = P_t + A_t(G_t - P_t)/p_t$. The learned construction extends this update by choosing how much each guidance–feedback rule contributes. For a fixed prediction P with finite residual second moment and a positive price κ_t, querying with probability p gives conditional error–cost $(p^{-1} - 1)\mathbb{E}_t\|G_t - P\|^2 + \kappa_t p$. If the residual second moment were known, its minimizing probability over $[\varepsilon, 1]$ would be

$$p^\star = \min\left\{1, \max\left\{\varepsilon, \sqrt{\frac{\mathbb{E}_t\|G_t - P\|^2}{\kappa_t}}\right\}\right\}.$$

At a fixed price, a smaller residual second moment warrants no greater query probability. This residual-based allocation principle is established in active inference (Zrnic & Candes, 2024). SRAO learns the rule weights from acquired references.

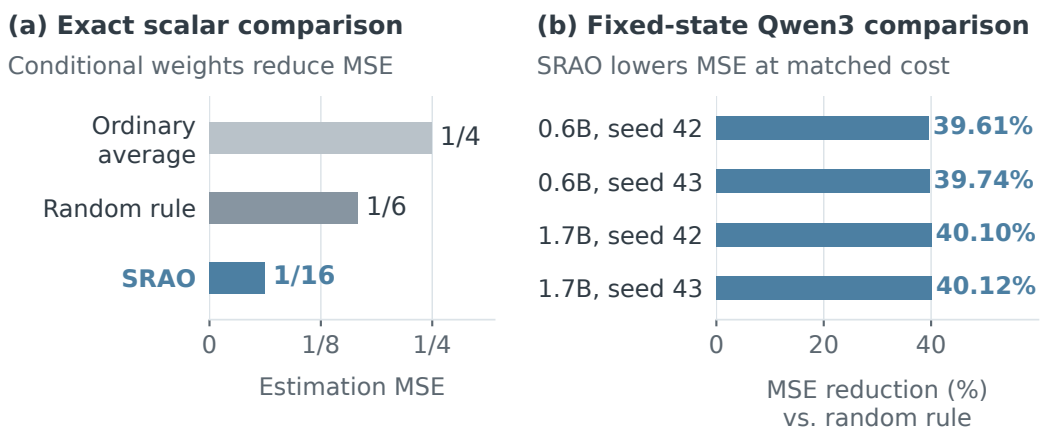


Figure 2: **Conditional averaging reduces estimation error at matched reference cost.** (a) For reference zero, predictions $(0, 1)$, equal prior weights and query probabilities $(1/4, 3/4)$, residual-corrected ordinary averaging, random-rule execution and SRAO have MSE $1/4$, $1/6$ and $1/16$, respectively. All three use the same expected query rate $1/2$. (b) At fixed Qwen3 states, SRAO reduces mean squared estimation error relative to random-rule execution at the same expected query rate $49/80$. Each row uses eight questions with eight pre-generated trajectories per question. These LLM comparisons use common predictions and fixed weights.

3.2 Learning from the same feedback

When $\bar{p}_t < 1$, the residual $R_{0,t} = G_t - P_{0,t}$ recovers every rule's residual as $G_t - P_{t,a} = R_{0,t} + P_{0,t} - P_{t,a}$. When $\bar{p}_t = 1$, the reference is always queried and these residuals are computed directly. For a fixed reference G_t, a rule's mean squared gradient-estimation error over its query coin, its importance-weighted error estimate and its expected weighted feedback cost are

$$h_{t,a}(G_t) = (p_{t,a}^{-1} - 1)\|G_t - P_{t,a}\|^2, \qquad e_t(a) = \frac{A_t}{\bar{p}_t} h_{t,a}(G_t), \qquad m_t(a) = \kappa_t p_{t,a}. \tag{2}$$

On an unqueried round, set $e_t = 0$ without evaluating G_t. **One acquired reference informs every rule.** The expensive observation is shared, while differences among the available predictions recover the rule-specific residuals. These importance-weighted errors let queried rounds inform future choices without additional reference queries. Algorithm 1 adapts AdaHedge's mixability-gap learning rate and exponential weights (de Rooij et al., 2014), incorporating the currently known feedback costs before the query. Only residual feedback determines the learning rate. For $\rho < \infty$, define $\Phi_\rho(z) = -\rho^{-1} \log(K^{-1} \sum_a e^{-\rho z_a})$ and set $\Phi_\infty(z) = \min_a z_a$. At infinite learning rate, the weights are uniform over *all current minimizers* of B_t. No rule is permanently discarded. Later feedback can make any rule a minimizer.

4 Theory and mechanism studies

The central question is whether SRAO's estimation gain can pay for learning how to combine guidance and acquisition. A sharp variance relation gives a horizon-independent expected excess bound under the stated conditions, and an optimization consequence separates acquisition error from reference noise and bias. Controlled scalar and gradient studies test the predicted gain.

4.1 Estimation and learning guarantees

We use the common-reference protocol of Section 3 and Euclidean norms. The analysis first identifies the estimation gain, then shows how it controls the variance cost of learning.

Fix one round and one realized G, and suppress t. Let $h_a = (p_a^{-1} - 1)\|G - P_a\|^2$ and $h_0 = (\bar{p}^{-1} - 1)\|G - P_0\|^2$. These are the mean squared gradient-estimation errors over the query decision,

Algorithm 1 SRAO: reference correction and acquisition learning

Require: $T \geq 1$, $K \geq 2$, $0 < \varepsilon \leq 1$, a declared state-update rule

Initialize x_1, $\widehat{L}_0 = 0$, $\Delta_0 = 0$
for $t = 1, \ldots, T$ **do**
  Fix common y_t, all $(P_{t,a}, p_{t,a})$, and κ_t before the query
  $m_t(a) = \kappa_t p_{t,a}$, $B_t = \widehat{L}_{t-1} + m_t$
  $\rho_t = \log K / \Delta_{t-1}$, using $\rho_t = \infty$ when $\Delta_{t-1} = 0$
  $r_{t,a} \propto \exp(-\rho_t B_t(a))$, compute $\bar{p}_t, P_{0,t}$ as in (1)
  Draw $A_t \sim \text{Bernoulli}(\bar{p}_t)$, query one G_t only if $A_t = 1$
  Form g_t, e_t by (1)–(2)
  Apply the declared state-update rule to g_t to obtain x_{t+1}
  $\delta_t = \langle r_t, e_t \rangle - [\Phi_{\rho_t}(B_t + e_t) - \Phi_{\rho_t}(B_t)]$
  $\Delta_t = \Delta_{t-1} + \delta_t$, $\widehat{L}_t = \widehat{L}_{t-1} + m_t + e_t$
end for
Return x_{T+1} for the fixed task evaluation

conditional on G, for feedback rule a and SRAO's combined estimate, respectively. With $d_a = G - P_a$, $D = G - P_0$, $u_a = (1 - p_a)d_a/p_a$ and $U = (1 - \bar{p})D/\bar{p}$, conditional projection gives

$$J := \sum_a r_a h_a - h_0 = \sum_a r_a J_a \geq 0, \qquad J_a = (1 - p_a)\|d_a - D\|^2 + p_a\|u_a - U\|^2. \tag{3}$$

The reduction J includes differences between rule estimates in both query outcomes. At $\bar{p} = 1$, set $h_0 = J = 0$. The conditional weighting is essential: ordinary averaging can be worse than sampling a rule, as Figure 2(a) shows. The next theorem connects the amount of error removed to the variation among the rules' errors.

Theorem 4.1 (Sharp bound relating rule errors and their reduction)**.** *Let $M = \max_a \|G - P_a\|^2$. For every finite set of feedback rules with $p_a \in [\varepsilon, 1]$,*

$$\text{Var}_{a \sim r}(h_a) + J^2 \leq M \sum_a \frac{r_a J_a}{p_a^2} \leq \frac{M}{\varepsilon^2} J. \tag{4}$$

At $\bar{p} = 1$, the weighted middle term is defined as zero. The coefficient ε^{-2} is uniformly sharp for each $0 < \varepsilon < 1$, even for two scalar feedback rules. For $0 < \varepsilon < 1$, no finite coefficient depending only on ε allows M to be replaced by $\max_a h_a$.

Proof idea. The identity $h_a - h_0 = \langle d_a - D, u_a \rangle + \langle D, u_a - U \rangle$ and weighted Cauchy–Schwarz give $(h_a - h_0)^2 \leq M J_a / p_a^2$, since D is a convex combination of the d_a. Averaging yields variance plus squared error reduction. Two-rule perturbations attain the floor coefficient in the limit. The complete proof and the counterexample for replacing the residual scale by the largest rule error appear in Appendix A. □

A common gradient prediction admits the smaller coefficient M/ε rather than M/ε^2 (equation (18)). Allowing different predictions therefore changes the sharp dependence on the query floor. This sharpness concerns the one-round inequality. It does not establish the optimal query-floor dependence of the online bound. The bridge applies to any predictable finite library sharing a reference, including cached gradients and inexpensive surrogate gradients. It therefore separates the statistical benefit of combining guidance from the particular model or procedure used to construct it.

Consider Algorithm 1 under the common-reference and pre-query requirements of Section 3, with deterministic $T \geq 1$, $K \geq 2$ and a fixed query floor $0 < \varepsilon \leq 1$. Assume the pathwise bound $M_t \leq V < \infty$ over every rule and round, and $\mathbb{E} \sum_{t \leq T} \kappa_t < \infty$. Writing $\mathbb{E}_t = \mathbb{E}[\cdot \mid \mathcal{H}_t]$, the expected comparison in Theorem A.5 gives

$$\mathbb{E} \sum_{t \leq T} [\mathbb{E}_t h_{t,0} + \kappa_t \bar{p}_t] \leq \min_a \mathbb{E} \sum_{t \leq T} [\mathbb{E}_t h_{t,a} + \kappa_t p_{t,a}] + \frac{V \log K}{\varepsilon^3} + \left(2 + \frac{4 \log K}{3}\right) \frac{V(1 - \varepsilon)}{\varepsilon^2}. \tag{5}$$

The comparator is the best fixed rule identity in expectation, evaluated on the learner's history, where its prediction and query probability can depend on the current context. For fixed K, ε and V, **the**

expected cumulative excess does not grow with the training horizon. The algorithm need not know V. Reference cost covers the queried computation, with guidance construction, rollouts and controller execution accounted for separately.

The sharp bridge links the variance of importance-weighted rule errors to the cumulative error removed by combining their estimates over both query outcomes. The variance-sensitive AdaHedge penalty (de Rooij et al., 2014) grows with the square root of this removed error, whereas the estimation benefit subtracts it directly. The cancellation has the form

$$2\sqrt{az} - z \le a \qquad (a, z \ge 0),$$

where z is the cumulative importance-weighted error reduction and a is the residual scale multiplied by $\log K$. The negative term is the estimator's accumulated advantage over sampling a rule. Thus the proof uses an actual improvement in the executed estimator to pay for learning its weights. This controls the variance-sensitive part of the learning penalty. The remaining range term gives the second term in the explicit remainder above. Incorporating known current costs before the query avoids an additional term for their range. Appendix A.2 gives the complete proof.

The result also separates the errors entering a general optimization update. Let $\mu_t = \mathbb{E}_t G_t$, $s_t = \mathbb{E}_t\|G_t - \mu_t\|^2$ and $\xi_t = \mu_t - \nabla f(x_t)$. Conditional unbiasedness gives

$$\mathbb{E}_t\|g_t - \nabla f(x_t)\|^2 = \|\xi_t\|^2 + s_t + \mathbb{E}_t h_{t,0}. \tag{6}$$

Selective acquisition controls an additive error term, separate from reference noise and mean mismatch. This gives a concrete optimization consequence. Consider $x_{t+1} = x_t - \eta g_t$ on a fixed L-smooth objective $f \ge f_*$, with $0 < \eta \le 1/L$. Let $\mathcal{O}_T^{\mathrm{J}}$ denote the comparator in (5) and C its explicit additive remainder. With the envelope and moments of Corollary A.9,

$$\begin{aligned}\frac{1}{T}\mathbb{E}\sum_{t\le T}[\|\nabla f(x_t)\|^2 + \kappa_t A_t] \le \frac{2(\mathbb{E}f(x_1) - f_*)}{\eta T} + \frac{\mathcal{O}_T^{\mathrm{J}} + C}{T} \\ + \frac{1}{T}\mathbb{E}\sum_{t\le T}[\|\xi_t\|^2 + s_t].\end{aligned} \tag{7}$$

The bound makes the role of SRAO explicit: **learning the rule weights adds at most C/T beyond the fixed-rule error–cost comparator**. The fixed-rule error–cost comparator and the reference's bias and noise remain. In particular, a small learning penalty does not require a perfect reference, but it does not remove an imperfect reference's optimization error. The full proof in Appendix A.3 retains displacement between the update and reference evaluation points.

Norm clipping also preserves a comparison to the same clipped reference, with the original comparator defined through raw estimation error (Corollary A.10). It does not require unbiasedness after clipping. The applications test how the resulting updates affect task quality.

4.2 How guidance and acquisition improve estimation

Combining rules converts the same feedback budget into a better estimate. Figure 2(a) isolates the conditional weighting in a scalar example. Table 1(a) applies the learning rule to three 1,024-step scalar sequences and compares the resulting execution objectives. SRAO reduces cumulative conditional estimation error plus reference cost relative to sampling the same learned rules in every construction. With the smaller query floor, the reduction is **46.55**, enough to improve on full feedback's objective of 512. In the moderate case, the observed excess over the best fixed rule is 1.40, covered by a potential certificate of 2.27. The uniform expected allowance is 18.79, only 3.67% of the full-feedback scale. Thus the learning guarantee can have an informative finite-horizon magnitude. The outlier case retains the combination gain but exceeds full feedback, and its maximum-based allowance becomes loose. Appendix B.1 gives the full certificate comparison.

The same mechanism persists in language-model gradients. We hold states, predictions and reference gradients fixed for Qwen3-0.6B and Qwen3-1.7B, using two seeds and eight GSM8K training questions per model. Exact enumeration of eight generated trajectories and both query outcomes removes query-mask sampling error. Figure 2(b) shows **about 40% lower MSE at identical expected reference cost** in all four model–seed configurations. Uniform weights give a further sensitivity check, with reductions of about 45%. This reduction also persists on the 11 mixed-reward groups,

Table 1: **Combination gains appear in scalar learning and heterogeneous-gradient histories.** (a) Conditional squared error plus reference cost over 1,024 steps. Both estimators use the same realized SRAO history, weights and expected query cost. Full feedback costs 512, and the outlier remains worse. (b) The no-query branch contributes to observed importance-weighted gain in all three distillation histories. Geometry is measured before clipping on 96 adaptive rounds per history.

(a) Scalar learning

Construction	Random rule	SRAO	Error removed
Moderate query floor	506.63	504.95	1.68
Small query floor	541.00	494.45	**46.55**
Single outlier	517.80	516.12	1.68

(b) Heterogeneous guidance during distillation

Seed	Observed gain	No-query share	Positive queried rounds
42	0.45	15.58%	61/61
43	0.66	14.35%	60/60
44	0.58	14.42%	60/60

where reward differences produce nonzero advantages. The gain is therefore present beyond the equal-reward groups with zero outcome gradients. These gains improve the combination of available rules even when their guidance is poorly aligned with the reference (Appendix B.1).

Distinct predictions contribute beyond mixing query rates. Table 1(b) decomposes gain in three distillation histories with zero and cached predictions. The no-query branch contributes about 15% in each history and is positive on all 181 queried adaptive rounds. This is the component that disappears when every rule uses the same prediction. It connects the two-branch identity in (3) to heterogeneous guidance during training. The learned histories still have positive importance-weighted excess over the best fixed observed rule, so this estimation gain does not establish superior online allocation at the measured horizon (Appendix B.2).

Guidance also identifies where references are valuable. A frozen allocation rule fitted to guidance length and gradient norm reduces estimation variance on all three held-out 100-step histories and two outcome-only anchors, one exploratory. Each comparison matches a constant rate to the same expected reference-token cost on a fully observed trajectory, and uses zero prediction to isolate allocation. The gain comes from concentrating feedback on larger gradient energy. An online mixture loses the fixed rule's advantage on one anchor, although combining its learned rules still improves estimation. These results identify useful acquisition information, while the cost of constructing guidance and the quality of online adaptation determine its eventual training benefit (Appendix B.1).

5 Learning applications

The applications test whether coordinating guidance and feedback improves the quality or efficiency of training. In on-policy distillation, SRAO retains similar observed mean accuracy with fewer reference gradients. In agentic question answering, coordinating reflection guidance with outcome feedback improves mean task quality. Table 3 specifies the correction operator that implements the shared principle in each setting.

5.1 Accuracy and reference computation in distillation

On-policy distillation tests whether SRAO can retain the benefit of teacher guidance while computing fewer fresh reference gradients. We instantiate SRAO with the objective of OPSD (Zhao et al., 2026). The most recent reference gradient supplies cached guidance, and a new reference corrects it through equation (1). Each *reference-gradient batch* comprises teacher inference and a backward pass at the current student state. Student generation continues at every step.

We train Qwen3-1.7B with rank-64 LoRA for 100 steps and 32 responses per step. Full feedback acquires every reference. SRAO uses fixed acquisition at $p = 1/2$ or learns among eight guidance–

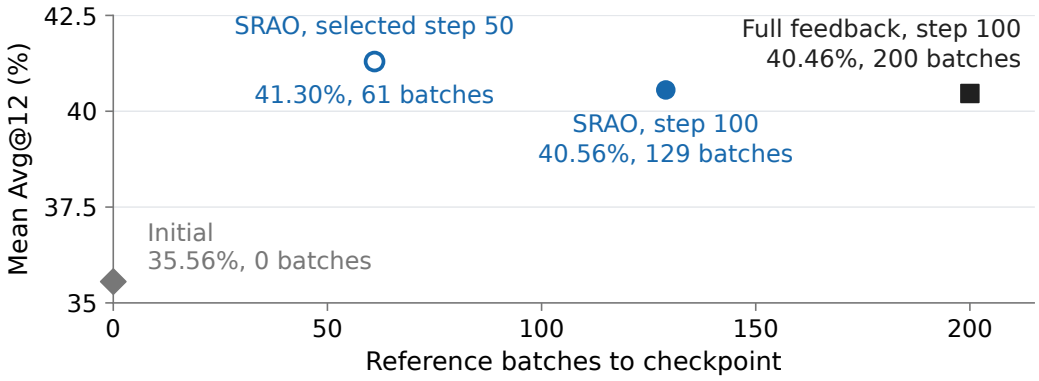


Figure 3: **SRAO retains the accuracy gain from distillation with fewer references.** Step-100 points average seeds 42 and 44. The open marker shows SRAO's step-50 checkpoint chosen by a common screen, which selects step 100 for full feedback. Reference batches are summed through the displayed checkpoints, including calibration. Initial is one evaluation of the shared starting model. The selected-checkpoint comparison includes the screening questions.

acquisition rules. Both include four calibration steps and share full feedback's AdamW update and clipping. We evaluate all 30 questions in AIME 2024, AIME 2025 and HMMT February 2025, with 12 answers per question and a common A800 profile. Avg@12 equally weights answer correctness across tasks and training seeds. Appendix B.2 gives the full setting.

Figure 3 shows that **selective reference acquisition retains the accuracy gain from distillation**. The starting model scores 35.56%. At step 100, SRAO reaches 40.56% across seeds 42 and 44, compared with 40.46% for full feedback, while using 129 rather than 200 reference batches. Thus SRAO gains 5.00 percentage points over the starting model with **35.5% fewer reference-gradient batches**. A third-seed replication reaches 41.20% versus 40.19%, using 64 rather than 100 batches. Task effects differ, including a 1.81-point reduction on AIME 2025 in the two-seed comparison. The saving concerns reference computation. Student completion tokens increase by 0.12%, and teacher-token totals are unavailable.

Earlier checkpoints further improve the observed accuracy–reference trade-off. We screen steps 25, 50, 75 and 100 on a shared subset, selecting one checkpoint per method across both seeds and all tasks. The selected SRAO checkpoint at step 50 reaches **41.30% with 61 reference batches**, compared with 40.46% and 200 batches for full feedback at its selected step 100. The 69.5% reduction covers the checkpoint prefixes. All runs continued to step 100. The full evaluation includes screening questions, making this a benchmark-informed comparison. Appendix B.2 gives the screen and the task-level results, including a 0.83-point decrease on HMMT.

Cached guidance is effective with either acquisition setting. At seed 44, fixed acquisition achieves 41.48% with 61 batches, compared with 40.56% and 64 batches for learned acquisition. Zero guidance gives 40.09% with 60 batches. This comparison supports reference correction, without establishing an advantage from learning acquisition.

Published OPSD reports 43.40% after selection over checkpoints, 2.10 points above SRAO's shared-checkpoint mean. Its unmeasured reference count precludes a matched accuracy–cost comparison. Appendix B.2 compares the published baselines and protocols.

5.2 Guidance and outcomes in agentic question answering

Agent learning tests SRAO with feedback available at every step. Reflection and backtracking guide intermediate decisions, while terminal rewards indicate whether the resulting trajectory succeeds. Auxiliary Adam steps first produce a candidate policy. Rescoring the same trajectories under that policy gives an outcome-driven update. The accepted update combines the outcome displacement with the guidance displacement after removing its opposing component.

Table 2: **SRAO improves mean answer quality at both scales, and the 3B component study supports coordination.** Answer-token F1 (%, ↑), with component drops in percentage points. Other five averages the five tasks outside the two training domains. All seven tasks enter SRAO configuration selection. Training data and computation differ across systems. Complete baseline comparisons and per-task results appear in Appendix B.3.

	3B		7B	
Method	All seven	Other five	All seven	Other five
Base (Qwen Team, 2025)	18.33	16.05	23.34	22.27
CoT (Wei et al., 2022)	19.61	18.92	27.71	28.26
Search-R1 (Jin et al., 2025)	30.16	28.80	34.08	32.93
SRAO	**46.15**	**43.77**	**49.21**	**47.35**

3B component setting	All seven F1	Drop from SRAO
SRAO	**46.15**	0.00
Without reflection	30.84	15.31
Without backtracking	33.09	13.06
Without fast distillation	40.51	5.64
Without direct outcome optimization	39.11	7.04
Without look-ahead coordination	41.21	4.94

We train Qwen2.5-Instruct models at 3B and 7B on HotpotQA and 2Wiki, then evaluate answer-token F1 on these domains and five additional QA tasks. Retrieval methods share the Wikipedia snapshot, retriever, inference budgets and answer normalizer. Training data and computation differ across the compared systems. The fixed-seed SRAO configuration uses three auxiliary steps and projected-guidance weight one, selected by the same seven-task mean reported here.

Table 2 compares SRAO with Base, CoT and Search-R1, the strongest listed seven-task-mean baseline at both scales. The complete comparison in Appendix B.3 shows SRAO's highest displayed F1 on every task at both scales. Its advantage also extends across the five tasks outside the training domains: the Other five mean exceeds Search-R1 by **14.97 points at 3B and 14.43 at 7B**, computed before rounding. These tasks contribute to configuration selection, so the result demonstrates broad task quality for the selected system.

The component panel connects that quality to the training design. Removing reflection, backtracking, fast distillation or direct outcome optimization lowers the seven-task mean. Retaining both training signals while removing look-ahead coordination produces a **4.94-point decrease**, supporting coordination beyond the mere presence of guidance and outcomes. This intervention jointly removes rescoring, conflict projection and their associated training work.

The mean gain coexists with task-specific trade-offs. Without coordination, Natural Questions improves from 55.37 to 55.94. Removing fast distillation improves Natural Questions and FRAMES while reducing the seven-task mean.

6 Conclusion

SRAO uses corrective feedback to coordinate guidance use and reference acquisition. Its sharp variance relation explains how the estimation gain from combining rules can offset the cost of learning them. The optimization consequence separates selective acquisition from reference noise and bias. Scalar and gradient studies test these mechanisms, while distillation and agentic learning test reference efficiency and update coordination. The applications share one principle: *feedback should improve both the update it corrects and the subsequent use of guidance*. This perspective can inform learning systems that combine cached computations, teacher supervision or auxiliary objectives with feedback of different costs and granularity.

AI USE STATEMENT

In this work, AI tools were used for language polishing, figure design and generation, coding support. Specifically, large language models assisted with improving manuscript clarity and readability, suggesting alternative phrasings, developing and debugging code. All AI-assisted text, code, mathematical statements, derivations, proofs, and figures were independently reviewed, verified, and revised by the authors before inclusion. The research questions, core methodology, scientific contributions, experimental design, interpretation of results, and final decisions were determined by the authors. The authors take full responsibility for the correctness, originality, and integrity of the final paper.

ETHICS STATEMENT

This work studies feedback-efficient optimization through controlled estimation experiments, on-policy distillation, and agentic question answering. The experiments use publicly available language models, mathematical reasoning datasets, question-answering benchmarks, and a Wikipedia-based retrieval corpus. The study does not involve human participants or animal subjects, and no private user data were collected. Models and datasets were used in accordance with their applicable licenses and terms of use. The agentic experiments are confined to benchmark question answering and retrieval, rather than deployment in consequential real-world settings. Improved benchmark performance does not establish factual reliability, safety, or suitability for high-stakes applications; the resulting models may retain biases and generate incorrect or misleading answers.

REPRODUCIBILITY STATEMENT

The selective-acquisition protocol, estimator, and adaptive rule-weight update are specified in Section 3 and Algorithm 1. Section 4 states the theoretical guarantees and their assumptions, with detailed proofs in Appendix A. Appendix B.1 describes the scalar mechanism studies and controlled gradient experiments using Qwen3-0.6B and Qwen3-1.7B on GSM8K. Appendix B.2 specifies the Qwen3-1.7B on-policy distillation experiments, including the OpenThoughts mathematics training data released with OPSD, guidance–acquisition rule library, calibration procedure, hyperparameters, random seeds, reference-cost accounting, and evaluation on AIME 2024, AIME 2025, and HMMT February 2025. Appendix B.3 describes the Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct agentic experiments, including training-data construction, optimization and coordination operators, retrieval settings, and evaluation on HotpotQA, 2WikiMultiHopQA, Bamboogle, FRAMES, MuSiQue, Natural Questions, and TriviaQA. The reported protocols distinguish fixed-step results from benchmark-informed checkpoint and configuration selection, and separate reference-computation savings from other training and evaluation costs. The selective-gradient guarantees apply under the stated common-reference assumptions; the agentic coordination mechanism is evaluated empirically as a separate application.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

Ruicheng Ao, Hongyu Chen, Haoyang Liu, David Simchi-Levi, and Will Wei Sun. PPI-SVRG: Unifying prediction-powered inference and variance reduction for semi-supervised optimization. arXiv preprint arXiv:2601.21470, 2026.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002. doi: 10.1137/S0097539701398375.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 49–56, 2009.

Zalan Borsos, Andreas Krause, and Kfir Y. Levy. Online variance reduction for stochastic optimization. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 324–357. PMLR, 2018.

Zalán Borsos, Sebastian Curi, Kfir Yehuda Levy, and Andreas Krause. Online variance reduction with mixtures. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 705–714. PMLR, 2019.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. ReSearch: Learning to reason with search for LLMs via reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2858. arXiv:2503.19470.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 794–803. PMLR, 2018.

Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(37):1281–1316, 2014.

Guanting Dong, Licheng Bao, Zhongyuan Wang, Kangzhi Zhao, Xiaoxi Li, Jiajie Jin, Jinghan Yang, Hangyu Mao, Fuzheng Zhang, Kun Gai, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. Agentic entropy-balanced policy optimization, 2025a. arXiv:2510.14545.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. Agentic reinforced policy optimization, 2025b. arXiv:2507.19849.

Yunshu Du, Wojciech M. Czarnecki, Siddhant M. Jayakumar, Mehrdad Farajtabar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224*, 2018.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for effective hybrid evaluation of language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 111489–111514. Curran Associates, Inc., 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Kaixuan Ji, Jiafan He, and Quanquan Gu. Reinforcement learning from human feedback with active queries. *Transactions on Machine Learning Research*, 2025.

Xinke Jiang, Yue Fang, Rihong Qiu, Haoyu Zhang, Yongxin Xu, Hao Chen, Wentao Zhang, Ruizhe Zhang, Yuchen Fang, Xinyu Ma, Xu Chu, Junfeng Zhao, and Yasha Wang. TC-RAG: Turing-complete RAG's case study on medical LLM systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11400–11426, 2025. doi: 10.18653/v1/2025.acl-long.558.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Ö. Arık, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*, 2025.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017. doi: 10.18653/v1/P17-1147.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020.

Jyrki Kivinen and Manfred K. Warmuth. Averaging expert predictions. In Paul Fischer and Hans Ulrich Simon (eds.), *Computational Learning Theory*, volume 1572 of *Lecture Notes in Computer Science*, pp. 153–167. Springer Berlin Heidelberg, 1999.

Wouter M. Koolen, Peter Grünwald, and Tim van Erven. Combining adversarial guarantees and stochastic fast rates in online learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4745–4759, 2025. doi: 10.18653/v1/2025.naacl-long.243.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, 2021. arXiv:2110.14048.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36577–36590. PMLR, 2024.

Hongseok Namkoong, Aman Sinha, Steve Yadlowsky, and John C. Duchi. Adaptive sampling probabilities for non-smooth optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2574–2583. PMLR, 2017.

Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. Optimal model management for multifidelity Monte Carlo estimation. *SIAM Journal on Scientific Computing*, 38(5):A3163–A3194, 2016.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.

Qwen Team. Qwen2.5 technical report, 2025. arXiv:2412.15115v2.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pp. 993–1019. PMLR, 2013.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

Noa Shoham, Ron Dorfman, Shalev Shaer, Kfir Y. Levy, and Yaniv Romano. Prediction-powered semi-supervised learning with online power tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3632. arXiv:2510.22586.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, 2023. doi: 10.18653/v1/2023.acl-long.557.

Dirk van der Hoeven, Nikita Zhivotovskiy, and Nicolò Cesa-Bianchi. A regret-variance trade-off in online learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 35188–35200. Curran Associates, Inc., 2022.

Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16(54):1793–1861, 2015.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report, 2025. arXiv:2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 2020. arXiv:2001.06782.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. arXiv preprint arXiv:2601.18734v3, 2026.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. MEM1: Learning to synergize memory and reasoning for efficient long-horizon agents, 2025. arXiv:2506.15841.

Tijana Zrnic and Emmanuel Candes. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62993–63010. PMLR, 2024.

Guide to the appendix. Appendix A explains how SRAO's estimation gain controls the statistical cost of learning guidance–acquisition rules, with complete proofs and an optimization analysis. Appendix B connects these results to controlled mechanism studies and the two learning applications. It details the accuracy–reference-cost comparison in distillation and the guidance–outcome component studies in agentic question answering. Appendix C positions SRAO within prediction-assisted learning, gradient coordination, adaptive acquisition, and online learning.

A Mathematical analysis

This appendix states the observed certificate and the full expected comparison underlying the consequence in Section 4.1. It also records the conditional optimization and clipping consequences. The common-reference information structure and Algorithm 1 are unchanged throughout. The proof chain is as follows. Proposition A.1 derives the conditional estimate, and Lemma A.2 computes its exact error reduction. Theorem A.3 links this reduction to rule-loss variance. Lemma A.6 transfers that relation to selectively observed errors, and Theorem A.7 proves the learning bound. Corollary A.8 supplies its integrable certificate hierarchy. The clipping and stationarity consequences then use, respectively, nonexpansion and smooth descent.

A.1 Estimation and sharp geometry

SRAO aggregates a library of predictable feedback rules. Every gradient prediction is available before the query, and every feedback rule refers to the same potential reference direction at one common predictable evaluation point. Its feedback cost charges only the complete computation of that reference. Constructing gradient predictions and their features is a separate cost. The results do not compare different proposal states, uncomputed gradient predictions, or separately rerun training trajectories.

Condition on a pre-query history $\mathcal{H}$. A finite library supplies vectors $P_a \in \mathbb{R}^d$ and query probabilities $p_a \in [\varepsilon, 1]$, where $0 < \varepsilon \le 1$. Fix weights $r_a \ge 0$ with $\sum_a r_a = 1$. For a latent comparison, draw $I \sim r$ and $A \mid I = a \sim \mathrm{Bernoulli}(p_a)$, and form

$$\widetilde{g} = P_I + \frac{A}{p_I}(G - P_I). \tag{8}$$

The common potential reference direction G is conditionally independent of (I, A) given $\mathcal{H}$. All gradient predictions and query probabilities are $\mathcal{H}$-measurable. Conditional second moments are assumed whenever a conditional variance or squared error is used. The pointwise inequalities below hold for every fixed finite reference value without a distributional bound on G. For ordinary vector conditional expectations, assume $\mathbb{E}[\|G\| + \sqrt{M}] < \infty$, where $M = \max_a \|G - P_a\|^2$. This condition makes the reference, predictions and both estimators integrable. The error-only inequalities do not require integrability of the absolute reference norm.

Proposition A.1 (Conditional gradient prediction and observable projection)**.** *Let $\bar{p} = \sum_a r_a p_a$. If $\bar{p} < 1$, define*

$$P_0 = \frac{\sum_a r_a(1 - p_a)P_a}{1 - \bar{p}}, \qquad g^{\mathrm{J}} = P_0 + \frac{A}{\bar{p}}(G - P_0). \tag{9}$$

Then $A \mid \mathcal{H} \sim \mathrm{Bernoulli}(\bar{p})$ and

$$g^{\mathrm{J}} = \mathbb{E}[\widetilde{g} \mid \mathcal{H}, A, AG] = \mathbb{E}[\widetilde{g} \mid \mathcal{H}, A, G]. \tag{10}$$

In particular, $\mathbb{E}[g^{\mathrm{J}} \mid \mathcal{H}, G] = G$. If $\bar{p} = 1$, every positive-weight feedback rule queries with probability one. Set $g^{\mathrm{J}} = G$ directly. No definition of P_0 by division by zero is needed.

Proof. On $A = 0$, the posterior probability of a feedback rule is $r_a(1 - p_a)/(1 - \bar{p})$ and $\widetilde{g} = P_I$, so its conditional mean is P_0. This mean is unchanged if G is additionally conditioned upon. On $A = 1$, independence gives posterior $r_a p_a/\bar{p}$ even after observing G. Therefore

$$\mathbb{E}[\widetilde{g} \mid \mathcal{H}, A = 1, G] = \frac{G}{\bar{p}} - \frac{\sum_a r_a(1 - p_a)P_a}{\bar{p}} = P_0 + \frac{G - P_0}{\bar{p}}.$$

The two branches prove both conditional-expectation statements. Averaging over the independent Bernoulli query decision at fixed G gives mean G. The case where feedback is always queried is immediate. □

Indeed, when $\bar{p} < 1$,

$$\mathbb{E}[\|g^{\mathrm{J}} - G\| \mid \mathcal{H}, G] = 2(1-\bar{p})\|G - P_0\| \le 2\sqrt{M},$$

and the sampled-rule counterpart is at most $2\sum_a r_a(1-p_a)\|G-P_a\| \le 2\sqrt{M}$. The all-query case is immediate. The observable sigma-field in (10) drops the latent feedback rule identity. SRAO can draw its query decision directly. Sampling I is only a comparison device. The combined gradient prediction is the *no-query posterior* average, not generally $\sum_a r_a P_a$.

For the next statements fix G and assume $\bar{p} < 1$. Put

$$\begin{aligned} d_a &= G - P_a, & D &= G - P_0, \\ u_a &= \frac{1-p_a}{p_a} d_a, & U &= \frac{1-\bar{p}}{\bar{p}} D, \\ h_a &= \frac{1-p_a}{p_a}\|d_a\|^2 = d_a \cdot u_a, & h_0 &= \frac{1-\bar{p}}{\bar{p}}\|D\|^2 = D \cdot U. \end{aligned} \tag{11}$$

The sampled-rule estimator's error about G is $-d_a$ on an unqueried round and u_a on a queried round. The corresponding SRAO errors are $-D$ and U.

Lemma A.2 (Error reduction across both query outcomes)**.** *Define*

$$J_a = (1-p_a)\|d_a - D\|^2 + p_a\|u_a - U\|^2, \qquad J = \sum_a r_a h_a - h_0. \tag{12}$$

Then

$$\begin{aligned} J &= \sum_a r_a J_a \ge 0, \\ \mathbb{E}[\|\widetilde{g} - G\|^2 \mid \mathcal{H}, G] &= \mathbb{E}[\|g^{\mathrm{J}} - G\|^2 \mid \mathcal{H}, G] + \mathbb{E}[\|\widetilde{g} - g^{\mathrm{J}}\|^2 \mid \mathcal{H}, G] \\ &= h_0 + J. \end{aligned} \tag{13}$$

Thus the error reduction includes both query and no-query branches and holds even conditional on the common random target's realized value.

Proof. The posterior means satisfy

$$\sum_a r_a(1-p_a)d_a = (1-\bar{p})D, \qquad \sum_a r_a p_a u_a = \bar{p}U.$$

Expanding the weighted squared distances in (12) therefore subtracts $(1-\bar{p})\|D\|^2 + \bar{p}\|U\|^2 = h_0$ from

$$\sum_a r_a\big[(1-p_a)\|d_a\|^2 + p_a\|u_a\|^2\big] = \sum_a r_a h_a.$$

The same branch calculation identifies these terms with the three conditional squared errors in (13). Alternatively, the second projection identity in (10) gives conditional orthogonality at fixed G. Integrating over G is valid when the relevant second moments are finite. □

Theorem A.3 (Sharp bound relating rule errors and their reduction)**.** *For $M = \max_a \|G - P_a\|^2$, the quantities above satisfy*

$$\mathrm{Var}_{a\sim r}(h_a) + J^2 \le M \sum_a \frac{r_a J_a}{p_a^2} \le \frac{M}{\varepsilon^2} J. \tag{14}$$

For each fixed $0 < \varepsilon < 1$, the coefficient ε^{-2} cannot be reduced uniformly over permitted sets of feedback rules, even with two scalar feedback rules. If $\bar{p} = 1$, set $h_0 = J = 0$ and define the weighted middle term in (14) *as zero. The weighted variance is zero. The case $\varepsilon = 1$ is entirely degenerate.*

Proof of the inequality. The identity $h_a - h_0 = (d_a - D)\cdot u_a + D \cdot (u_a - U)$ and weighted Cauchy–Schwarz give, for $p_a < 1$,

$$\begin{aligned} (h_a - h_0)^2 &\le J_a \left(\frac{\|u_a\|^2}{1-p_a} + \frac{\|D\|^2}{p_a} \right) \\ &= J_a \left(\frac{1-p_a}{p_a^2}\|d_a\|^2 + \frac{\|D\|^2}{p_a} \right) \le \frac{M}{p_a^2} J_a. \end{aligned} \tag{15}$$

The last step uses that D is a convex combination of the d_a, with weights $r_a(1-p_a)/(1-\bar{p})$, so $\|D\|^2 \le M$. If $p_a = 1$, then $u_a = 0$ and $J_a = \|U\|^2$. Directly, $(h_a - h_0)^2 = (D \cdot U)^2 \le M J_a$. Finally,

$$\sum_a r_a (h_a - h_0)^2 = \mathrm{Var}_{a\sim r}(h_a) + \left(\sum_a r_a h_a - h_0\right)^2 = \mathrm{Var}_{a\sim r}(h_a) + J^2.$$

Average (15), use $p_a \ge \varepsilon$, and apply Lemma A.2. The case where feedback is always queried and the case $M = 0$ each give zero weighted variance and error reduction. Sharpness is proved next. □

Proof of sharpness. Fix $0 < \varepsilon < 1$. Use two equally weighted scalar feedback rules, $G = 0$, and

$$d_1 = 1, \quad d_2 = 1 + s, \quad P_a = -d_a, \qquad p_1 = \varepsilon, \quad p_2 = \frac{\varepsilon(1+s)}{1+2\varepsilon s}. \tag{16}$$

For $\varepsilon \le 1/2$, take $s > 0$ tending to zero. For $\varepsilon > 1/2$, take $-(1-\varepsilon)/\varepsilon < s < 0$ tending to zero. In both cases $p_2 \in [\varepsilon, 1)$ and the actual minimum query probability is ε. By construction,

$$u_1 = (1-\varepsilon)/\varepsilon, \qquad u_2 = u_1 + s.$$

Writing $q = (p_1 + p_2)/2$, both branch-value differences equal s, so the two-point weighted-variance formula in (12) gives

$$\begin{aligned} \mathrm{Var}_r(h_a) &= \frac{s^2}{4}(\varepsilon^{-1} + s)^2, \qquad & M &= \max\{1, (1+s)^2\}, \\ J &= \frac{s^2}{4}\left(1 - \frac{(p_2 - p_1)^2}{4q(1-q)}\right). \end{aligned} \tag{17}$$

For example, the two branch contributions to J/s^2 are $(1-p_1)(1-p_2)/[2(2-p_1-p_2)]$ and $p_1 p_2/[2(p_1+p_2)]$, whose sum is the displayed factor. It follows that

$$\lim_{s\to 0} \frac{\mathrm{Var}_r(h_a)}{MJ} = \varepsilon^{-2}.$$

The same limit holds after adding J^2 to the numerator, since $J/M \to 0$. Thus neither version of the bridge admits a smaller uniform coefficient. At $\varepsilon = 1/2$ this family has equal query probabilities and different gradient predictions. □

If all $P_a = P$, write $M = \|G - P\|^2$. Then $P_0 = P$ and the stronger inequality for a common gradient prediction is

$$\mathrm{Var}_r(h_a) + J^2 \le \frac{M}{\varepsilon} J. \tag{18}$$

Indeed,

$$\mathbb{E}_r(1/p_a - 1/\bar{p})^2 = \mathbb{E}_r \frac{(p_a - \bar{p})^2}{p_a^2 \bar{p}^2} \le \frac{1}{\varepsilon}\mathbb{E}_r \frac{(p_a - \bar{p})^2}{p_a \bar{p}^2} = \frac{1}{\varepsilon}(\mathbb{E}_r p_a^{-1} - \bar{p}^{-1}).$$

Multiplying by M^2 proves (18). The distinction between ε^{-1} and ε^{-2} here concerns the geometry of the two theorem classes. It is not an empirical ranking of their controllers or an online minimax lower bound.

Proposition A.4 (The largest rule gradient-estimation error is not a sufficient scale)**.** *For any fixed $0 < \varepsilon < 1$, there is no finite coefficient depending only on ε such that $\mathrm{Var}_r(h_a) \le C(\varepsilon)(\max_a h_a)J$ for every permitted library.*

Proof. Use equal weights, $G = 0$, identical gradient predictions $P_1 = P_2 = -1$, and query probabilities $p_1 = 1 - \delta$, $p_2 = 1$, where $0 < \delta \le 1 - \varepsilon$. Then

$$\begin{aligned} h_1 &= \frac{\delta}{1-\delta}, \qquad & h_2 &= 0, \qquad & h_0 &= \frac{\delta}{2-\delta}, \\ J &= \frac{\delta^2}{2(1-\delta)(2-\delta)}, \quad & \mathrm{Var}_r(h_a) &= \frac{\delta^2}{4(1-\delta)^2}. \end{aligned} \tag{19}$$

Consequently

$$\frac{\mathrm{Var}_r(h_a)}{(\max_a h_a)J} = \frac{2-\delta}{2\delta} \longrightarrow \infty.$$

For a concrete failure of the replacement $M \mapsto \max_a h_a$ in (14), take $\varepsilon = 1/2$ and $\delta = 1/10$. The variance is $1/324$, while $J = 1/342$, $\max_a h_a = 1/9$, and the proposed replacement right side is $2/1539 < 1/324$. □

A.2 Learning from shared feedback

The analysis separates a certificate computed from queried feedback from an upper bound on its expectation. Write $b = \log K$, $\widehat{M}_t = A_t M_t/\bar{p}_t$, $\widehat{J}_t = A_t J_t/\bar{p}_t$, and define

$$\begin{aligned} H_T^{\mathrm{J}} &= \max_{t\le T} \widehat{M}_t/\varepsilon^2, & S_T^{\mathrm{J}} &= \max_{t\le T}[\max_a e_t(a) - \min_a e_t(a)], \\ B_T^{\mathrm{pot}} &= \Delta_T + \Delta_{T-1} - \sum_t \widehat{J}_t - \sum_t \zeta_t, & \mathcal{D}_T^{\mathrm{J}} &= \mathbb{E}[bH_T^{\mathrm{J}} + (2+4b/3)S_T^{\mathrm{J}}]. \end{aligned} \tag{20}$$

The nonnegative adjustment $\zeta_t = \Phi_{\rho_t}(B_t) - \Phi_{\rho_t}(\widehat{L}_{t-1}) - \langle r_t, m_t\rangle \ge 0$ accounts for using the known current feedback costs before the query. The potential argument gives $\sum_t \widehat{f}_t - \min_a \widehat{L}_{T,a} \le B_T^{\mathrm{pot}}$, where $\widehat{f}_t = \kappa_t \bar{p}_t + A_t h_{t,0}/\bar{p}_t$. This signed certificate uses only queried references and known costs.

The sharp bridge then bounds the variation in the observed errors: $Q_t = \mathrm{Var}_{r_t}(e_t) \le H_T^{\mathrm{J}} \widehat{J}_t$. The AdaHedge penalty is $2\sqrt{b\sum_t Q_t} + (2+4b/3)S_T^{\mathrm{J}}$, while combining the rules saves $\sum_t \widehat{J}_t$ in the estimated cumulative error. For $U = \sum_t \widehat{J}_t$, the resulting tradeoff satisfies $2\sqrt{bH_T^{\mathrm{J}}U} - U \le bH_T^{\mathrm{J}}$. The saved error offsets the variance-sensitive part of learning the rules. Consequently, $\mathbb{E}B_T^{\mathrm{pot}} \le \mathcal{D}_T^{\mathrm{J}}$, without an additional term for the range of weighted feedback costs.

Theorem A.5 (Gradient-estimation error and feedback cost against a fixed rule)**.** *Run Algorithm 1 for deterministic T. Assume the pathwise envelope $M_t \le V < \infty$ over all feedback rules and $\mathbb{E}\sum_{t\le T}\kappa_t < \infty$. With $\mathbb{E}_t = \mathbb{E}[\cdot \mid \mathcal{H}_t]$, define $\mathcal{O}_T^{\mathrm{J}} = \min_a \mathbb{E}\sum_{t\le T}[\mathbb{E}_t h_{t,a} + \kappa_t p_{t,a}]$. Then*

$$\begin{aligned} \mathbb{E}\sum_{t\le T}[\mathbb{E}_t h_{t,0} + \kappa_t \bar{p}_t] &\le \mathcal{O}_T^{\mathrm{J}} + \mathbb{E}B_T^{\mathrm{pot}} \le \mathcal{O}_T^{\mathrm{J}} + \mathcal{D}_T^{\mathrm{J}}, \\ \mathcal{D}_T^{\mathrm{J}} &\le \frac{V\log K}{\varepsilon^3} + \left(2 + \frac{4\log K}{3}\right)\frac{V(1-\varepsilon)}{\varepsilon^2}. \end{aligned} \tag{21}$$

The comparator is one fixed rule identity evaluated on the learner's encountered history. Its gradient prediction and query probability may change predictably.

Proof idea. The potential with known weighted feedback costs gives the observed certificate. Take expectations against each fixed rule identity before minimizing. The bridge and absorption bound its expected remainder. The envelope bounds $H_T^{\mathrm{J}} \le V/\varepsilon^3$ and $S_T^{\mathrm{J}} \le V(1-\varepsilon)/\varepsilon^2$. The complete argument follows in Theorem A.7, including the adaptive potential with known weighted feedback costs. □

The algorithm does not require V as an input. Its observed certificate B_T^{pot} bounds the realized importance-weighted comparison, while $\mathbb{E}B_T^{\mathrm{pot}}$ enters the expected comparison. Appendix A.2 states the sufficient integrability conditions and further certificates.

The payoff is an excess upper bound independent of the number of steps when K, ε, V are fixed. Thus $\mathcal{D}_T^{\mathrm{J}}/T$ is $O(1/T)$, although the signed remainder $\mathbb{E}B_T^{\mathrm{pot}}/T$ need not vanish. This conclusion requires a uniform residual bound across every rule. A poor prediction can enlarge that bound. The minimum is outside expectation and compares one rule identity on the learner's history, not switching identities or separately trained trajectories. A zero prediction is eligible with its own error and feedback cost. The work of constructing other predictions remains separately accounted.

Fix $K \ge 2$ feedback rule identities and a deterministic horizon T. Before each query, feedback rule a supplies $(P_{t,a}, p_{t,a})$ with $p_{t,a} \in [\varepsilon, 1]$. The current weighted feedback cost $\kappa_t \ge 0$ is known before

choosing the weights. All feedback rules refer to the same potential G_t at one common predictable evaluation point. The pre-query history $\mathcal{H}_t$ includes the past, all current gradient predictions, query probabilities and weighted feedback costs, and the controller's current weights and combined gradient prediction. With $\mathbb{E}_t = \mathbb{E}[\cdot \mid \mathcal{H}_t]$, require

$$A_t \mid \mathcal{H}_t \sim \text{Bernoulli}(\bar{p}_t), \qquad A_t \perp G_t \mid \mathcal{H}_t. \tag{22}$$

The law of G_t may depend on this encountered history. It need not be independent across rounds. The query condition rules out dependence on the current query coin. Internal controller randomness, if any, is included in the pre-query history before the query decision is drawn.

The SRAO update uses the residual-only AdaHedge recurrence with the gradient-estimation errors of its feedback rules. Put $b = \log K$, $\widehat{L}_0 = 0$, $\Delta_0 = 0$, and

$$\Phi_\rho(L) = -\rho^{-1} \log\left(K^{-1} \sum_a e^{-\rho L_a}\right), \qquad \Phi_\infty(L) = \min_a L_a. \tag{23}$$

At round t, form

$$\begin{aligned} m_t(a) &= \kappa_t p_{t,a}, & B_t &= \widehat{L}_{t-1} + m_t, \\ \rho_t &= b/\Delta_{t-1}, & r_{t,a} &\propto \exp[-\rho_t B_t(a)], \\ \bar{p}_t &= \sum_a r_{t,a} p_{t,a}, & P_{0,t} &= \frac{\sum_a r_{t,a}(1-p_{t,a})P_{t,a}}{1-\bar{p}_t}. \end{aligned} \tag{24}$$

When $\Delta_{t-1} = 0$, use $\rho_t = \infty$ and uniform weights on all current minimizers of B_t. The expression for $P_{0,t}$ is used only when $\bar{p}_t < 1$. Draw the query decision (22) and use $g_t = P_{0,t} + A_t(G_t - P_{0,t})/\bar{p}_t$, or $g_t = G_t$ when $\bar{p}_t = 1$.

For the common potential reference direction, define $h_{t,a}, h_{t,0}, J_t, M_t$ by the one-round definitions above, with $h_{t,0} = J_t = 0$ when feedback is always queried. After a query, one queried G_t evaluates every feedback rule. Without a query, set every residual score to zero. Thus

$$\begin{aligned} e_t(a) &= \frac{A_t}{\bar{p}_t} h_{t,a}, & \widehat{J}_t &= \frac{A_t}{\bar{p}_t} J_t, \quad \widehat{M}_t = \frac{A_t}{\bar{p}_t} M_t, \\ \widehat{f}_t &= \frac{A_t}{\bar{p}_t} h_{t,0} + \kappa_t \bar{p}_t. \end{aligned} \tag{25}$$

The products involving an unobserved reference mean the observable value zero on $A_t = 0$. The potential reference direction is used only to specify the probability model. An observed residual relative to $P_{0,t}$ also suffices, since adding the known difference $P_{0,t} - P_{t,a}$ reconstructs every feedback rule's residual. Update

$$\begin{aligned} \delta_t &= \langle r_t, e_t \rangle - [\Phi_{\rho_t}(B_t + e_t) - \Phi_{\rho_t}(B_t)], \\ \Delta_t &= \Delta_{t-1} + \delta_t, \qquad \widehat{L}_t = \widehat{L}_{t-1} + m_t + e_t. \end{aligned} \tag{26}$$

Neither a residual envelope nor the horizon is an input to this recurrence.

Lemma A.6 (Observed error reduction bound)**.** *On every finite realized path,*

$$\langle r_t, m_t + e_t \rangle - \widehat{f}_t = \widehat{J}_t \geq 0, \qquad Q_t := \text{Var}_{a \sim r_t}(e_t(a)) \leq \frac{\widehat{M}_t}{\varepsilon^2} \widehat{J}_t. \tag{27}$$

Proof. The feedback cost identity follows from $\sum_a r_{t,a} \kappa_t p_{t,a} = \kappa_t \bar{p}_t$. The residual identity is the definition of J_t. Multiply Theorem A.3 by $(A_t/\bar{p}_t)^2$ and discard its nonnegative square of the error reduction. The product of the two resulting scale factors is exactly $(\widehat{M}_t/\varepsilon^2)\widehat{J}_t$. This argument is pathwise and does not assume that M_t is predictable. When feedback is always queried, the weighted variance of the error estimates and the error reduction are both zero. □

Theorem A.7 (SRAO certificate against a fixed feedback rule)**.** *Under the preceding protocol, let*

$$\begin{aligned} H_T^{\text{J}} &= \max_{t \leq T} \widehat{M}_t/\varepsilon^2, & S_T^{\text{J}} &= \max_{t \leq T} [\max_a e_t(a) - \min_a e_t(a)], \\ \mathcal{D}_T^{\text{J}} &= \mathbb{E}\big[bH_T^{\text{J}} + (2 + 4b/3)S_T^{\text{J}}\big]. \end{aligned} \tag{28}$$

The nonnegative potential adjustment from the known current weighted feedback costs is $\zeta_t = \Phi_{\rho_t}(B_t) - \Phi_{\rho_t}(\widehat{L}_{t-1}) - \langle r_t, m_t \rangle$. *Pathwise,*

$$\sum_{t\le T} \widehat{f}_t - \min_a \sum_{t\le T} [m_t(a) + e_t(a)] \le bH_T^{\mathrm{J}} + (2 + 4b/3)S_T^{\mathrm{J}} - \sum_{t\le T} \zeta_t. \tag{29}$$

Assuming integrability of the displayed quantities, define

$$\mathcal{O}_T^{\mathrm{J}} = \min_a \mathbb{E} \sum_{t\le T} [\mathbb{E}_t h_{t,a} + \kappa_t p_{t,a}]. \tag{30}$$

Then

$$\mathbb{E} \sum_{t\le T} [\mathbb{E}_t h_{t,0} + \kappa_t \bar{p}_t] \le \mathcal{O}_T^{\mathrm{J}} + \mathcal{D}_T^{\mathrm{J}}. \tag{31}$$

In particular, if the deterministic pathwise envelope over all feedback rules $\max_a \|G_t - P_{t,a}\|^2 \le V < \infty$ *holds on every round and* $\mathbb{E} \sum_{t\le T} \kappa_t < \infty$, *then*

$$\mathcal{D}_T^{\mathrm{J}} \le \frac{V \log K}{\varepsilon^3} + \left(2 + \frac{4 \log K}{3}\right) \frac{V(1-\varepsilon)}{\varepsilon^2}. \tag{32}$$

The minimum is outside expectation and selects one fixed feedback rule identity on the learner's encountered history. Its gradient prediction and query probability may change predictably across rounds.

Proof. We first record the potential properties used below. For finite $\rho > 0$, let $v_a = \exp(-\rho L_a)/\sum_j \exp(-\rho L_j)$. For w in the probability simplex Δ_K, expansion of $\mathrm{KL}(w\|v) = \sum_{a:w_a>0} w_a \log(w_a/v_a)$ gives

$$\Phi_\rho(L) = \min_{w\in\Delta_K} \left\{ \langle w, L \rangle + \rho^{-1} \mathrm{KL}(w\|\mathrm{Unif}_K) \right\}.$$

The minimum is attained at v. Consequently Φ_ρ is concave in L, has gradient v, and is nonincreasing in ρ. Moreover, $\min_a L_a \le \Phi_\rho(L) \le \min_a L_a + b/\rho$. The limit at $\rho = \infty$ is $\min_a L_a$, with the algorithm's weights uniform on all minimizers.

We adapt the potential analysis of AdaHedge (de Rooij et al., 2014) to nonnegative residual-loss vectors e_t and predictable current cost vectors m_t under the stated recurrence. The optimistic potential identity, including its limit at infinite learning rate, is

$$\langle r_t, m_t + e_t \rangle = \Phi_{\rho_t}(\widehat{L}_t) - \Phi_{\rho_t}(\widehat{L}_{t-1}) + \delta_t - \zeta_t.$$

Concavity of Φ_ρ gives $0 \le \delta_t \le \omega_t$, where $\omega_t = \max_a e_t(a) - \min_a e_t(a)$, so ρ_t is nonincreasing. Since $\Phi_\rho(L)$ is nonincreasing in ρ, the additional terms from telescoping successive learning rates are nonpositive. The terminal excess over $\min_a \widehat{L}_{T,a}$ is at most $b/\rho_T = \Delta_{T-1}$. Hence

$$\sum_t \langle r_t, m_t + e_t \rangle - \min_a \widehat{L}_{T,a} \le \Delta_T + \Delta_{T-1} - \sum_t \zeta_t \le 2\Delta_T - \sum_t \zeta_t.$$

At finite learning rate, $\zeta_t = \rho_t^{-1} \mathrm{KL}(r_t\|w_t^0) \ge 0$, where $w_{t,a}^0 \propto e^{-\rho_t \widehat{L}_{t-1,a}}$. At infinite learning rate it equals $\langle r_t, \widehat{L}_{t-1} \rangle - \min_a \widehat{L}_{t-1,a} \ge 0$. We now derive the residual Bernstein bound, retaining the constants of the variance-sensitive AdaHedge analysis (de Rooij et al., 2014). At a finite rate $\rho > 0$, write $\ell = \langle r, e \rangle$, $X_a = \ell - e(a)$, $\omega = \max_a e(a) - \min_a e(a)$ and $Q = \sum_a r_a X_a^2$. The exponential weights at B give

$$\rho\delta = \log \sum_a r_a e^{\rho X_a}, \qquad \sum_a r_a X_a = 0, \qquad |X_a| \le \omega.$$

If $\omega = 0$, then $\delta = Q = 0$. When $\rho\omega < 3$, the Taylor series and $n! \ge 2\,3^{n-2}$ for $n \ge 2$ imply

$$\sum_a r_a e^{\rho X_a} \le 1 + \sum_{n\ge 2} \frac{\rho^n \omega^{n-2} Q}{n!} \le 1 + \frac{\rho^2 Q}{2(1 - \rho\omega/3)}.$$

Using $\log(1+z) \le z$ and rearranging yields

$$\frac{\delta}{\rho} \le \frac{Q}{2} + \frac{\omega\delta}{3}.$$

If $\rho\omega \ge 3$, the same inequality follows from $\delta \ge 0$ and $\rho^{-1} \le \omega/3$. At $\rho = \infty$, its left side is defined as zero and the right side is nonnegative. Thus it holds on every round. Together with $0 \le \delta_t \le \omega_t$, $\Delta_{t-1} = b/\rho_t$ and $\alpha = 1 + 2b/3$, it gives

$$\Delta_t^2 - \Delta_{t-1}^2 = 2b\delta_t/\rho_t + \delta_t^2 \le bQ_t + \alpha\omega_t\delta_t.$$

Summing and using $\omega_t \le S_T^{\mathrm{J}}$ gives $\Delta_T^2 \le b\sum_t Q_t + \alpha S_T^{\mathrm{J}}\Delta_T$. Solving this quadratic,

$$\Delta_T \le \frac{\alpha S_T^{\mathrm{J}} + \sqrt{\alpha^2 (S_T^{\mathrm{J}})^2 + 4b\sum_t Q_t}}{2} \le \alpha S_T^{\mathrm{J}} + \sqrt{b\sum_t Q_t}.$$

Therefore $2\Delta_T \le 2\sqrt{b\sum_t Q_t} + (2 + 4b/3)S_T^{\mathrm{J}}$. No geometry for a common gradient prediction has been used in these steps. Now put $U_T = \sum_t \widehat{J}_t$. Lemma A.6 gives $\sum_t Q_t \le H_T^{\mathrm{J}} U_T$ and subtracts U_T from the weighted loss. The inequality $2\sqrt{bH_T^{\mathrm{J}}U_T} - U_T \le bH_T^{\mathrm{J}}$ proves (29).

All rule outputs, weights and the combined gradient prediction are fixed in $\mathcal{H}_t$ before the query decision. Although $h_{t,0}$ is random through G_t, conditional independence implies

$$\mathbb{E}_t e_t(a) = \mathbb{E}_t h_{t,a}, \qquad \mathbb{E}_t[A_t h_{t,0}/\bar{p}_t] = \mathbb{E}_t h_{t,0}.$$

Apply (29) against each fixed identity first, take expectations, discard the nonnegative adjustment for known weighted feedback costs, and only then minimize over the identity. This proves (31) without replacing the comparator by a minimum inside expectation. Under the envelope, $\widehat{M}_t \le V/\varepsilon$ and

$$H_T^{\mathrm{J}} \le V/\varepsilon^3, \qquad 0 \le e_t(a) \le V(1-\varepsilon)/\varepsilon^2.$$

The latter also bounds S_T^{J}, proving (32). □

The envelope over all feedback rules is stronger than a bound only on the executed gradient prediction's residual. A poor feedback rule can enlarge it. Equation (32) is an upper bound for this controller. Sharpness of the one-step bridge does not imply that its ε^{-3} online dependence is minimax necessary. When gradient predictions coincide, the sharper geometry for a common gradient prediction (18) remains available.

The proof of Theorem A.7 supplies a tighter observable certificate before its maximum-based relaxation. The following restatement separates that quantity from the uniform expected bound. It does not change the controller or the sharp-bridge claim.

Corollary A.8 (Observable certificate hierarchy)**.** *For deterministic $T \ge 1$, put $b = \log K$, $c = 2 + 4b/3$, $U_T = \sum_{t\le T} \widehat{J}_t$ and $Z_T = \sum_{t\le T} \zeta_t$. Define*

$$B_T^{\mathrm{pot}} = \Delta_T + \Delta_{T-1} - U_T - Z_T,$$
$$B_T^{\mathrm{var}} = 2\sqrt{b\sum_{t\le T} Q_t} + cS_T^{\mathrm{J}} - U_T - Z_T,$$
$$B_T^{\mathrm{joint}} = bH_T^{\mathrm{J}} + cS_T^{\mathrm{J}} - Z_T.$$

All three are computable from queried reference directions and known weighted feedback costs. On every finite path of the mathematical recurrence,

$$\sum_{t\le T} \widehat{f}_t - \min_a \widehat{L}_{T,a} \le B_T^{\mathrm{pot}} \le B_T^{\mathrm{var}} \le B_T^{\mathrm{joint}}. \tag{33}$$

Under the query conditions of Theorem A.7 and integrability, the same fixed feedback rule comparator satisfies

$$\mathbb{E}\sum_{t\le T}(\mathbb{E}_t h_{t,0} + \kappa_t \bar{p}_t) \le \mathcal{O}_T^{\mathrm{J}} + \mathbb{E}B_T^{\mathrm{pot}} \le \mathcal{O}_T^{\mathrm{J}} + \mathcal{D}_T^{\mathrm{J}}. \tag{34}$$

A sufficient integrability condition is $\mathbb{E}\sum_{t\le T}(\kappa_t + M_t) < \infty$ for fixed K, T and $\varepsilon > 0$.

Proof. The potential bound with a varying learning rate in the preceding proof gives the first inequality in (33) after subtracting the cumulative error reduction U_T. Its Bernstein bound gives the second, and the observed bridge plus absorption gives the third. Apply the first bound against each fixed identity, take expectations using the two conditional inverse-probability-weighting identities, and only then minimize over that identity. Finally, $Z_T \geq 0$ implies $\mathbb{E}B_T^{\mathrm{joint}} \leq \mathcal{D}_T^{\mathrm{J}}$. For integrability, put $C_\varepsilon = (1-\varepsilon)/\varepsilon^2$. The nonnegative quantities $e_t(a), \delta_t, \widehat{J}_t$ are at most $C_\varepsilon M_t$, and $\zeta_t \leq \kappa_t$. Also $\sqrt{\sum_t Q_t} \leq C_\varepsilon \sum_t M_t$ and $H_T^{\mathrm{J}} \leq \varepsilon^{-3} \sum_t M_t$. Thus the stated condition controls every displayed expectation without requiring $\mathbb{E}Q_t < \infty$. □

Feedback rule outputs and cumulative losses can be random on the learner's encountered history. The comparator identity is fixed before the outcomes. A hindsight-selected random identity does not inherit this unbiasedness argument. This is not a comparison with a separately run training trajectory. In particular, a single observed B_T^{pot} cannot replace $\mathbb{E}B_T^{\mathrm{pot}}$ in (34). The deterministic envelope in (32) remains a sufficient conservative upper bound for this expectation.

B_T^{pot} is a *signed* comparison remainder. For example, at $T = 1$, gradient predictions $-1, +1$, both query probabilities $1/2$, $G = 0$ and $\kappa = 0$ give $B_1^{\mathrm{pot}} = -2A_1$ and $\mathbb{E}B_1^{\mathrm{pot}} = -1$. In general $\Delta_0 = 0$, $\zeta_1 = 0$ and $\widehat{R}_1 = B_1^{\mathrm{pot}} = \Delta_1 - \widehat{J}_1$. At $\bar{p}_t = 1$, the existing $h_{t,0} = \widehat{J}_t = 0$ convention avoids defining a no-query posterior. If every feedback rule always queries, $B_T^{\mathrm{pot}} = B_T^{\mathrm{var}} = 0$, whereas $B_T^{\mathrm{joint}} = bH_T^{\mathrm{J}}$ can remain positive.

A.3 Optimization and clipping

An optimization consequence requires specifying both the objective and the update. For a fixed smooth objective and a constant-step gradient update, the selective estimation error enters an exact decomposition alongside reference noise and mean mismatch. The following specialization retains all three terms.

Corollary A.9 (Conditional stationarity and feedback cost)**.** *Under the common-reference protocol and the integrability conditions of Corollary A.8, let $f : \mathbb{R}^d \to \mathbb{R}$ be a fixed L-smooth function, $L > 0$, bounded below by f_*. Specialize the declared update to $x_{t+1} = x_t - \eta g_t$, with constant $0 < \eta \leq 1/L$ and deterministic T. At the common predictable evaluation point y_t, assume*

$$\mu_t := \mathbb{E}_t G_t = \nabla f(y_t) + b_t, \qquad \xi_t = \mu_t - \nabla f(x_t), \qquad s_t = \mathbb{E}_t \|G_t - \mu_t\|^2, \tag{35}$$

where b_t is predictable reference bias. Assume $\mathbb{E}\sum_{t\leq T}(\|\xi_t\|^2 + s_t) < \infty$ and $\mathbb{E}|f(x_t)| < \infty$ for $1 \leq t \leq T+1$, in addition to the residual and cost integrability above. Select τ uniformly from $\{1, \ldots, T\}$ independently of the run. Then

$$\mathbb{E}_t \|g_t - \nabla f(x_t)\|^2 = \|\xi_t\|^2 + s_t + \mathbb{E}_t h_{t,0}, \tag{36}$$

and

$$\begin{aligned} \mathbb{E}\|\nabla f(x_\tau)\|^2 + \frac{1}{T}\mathbb{E}\sum_{t\leq T} \kappa_t A_t &\leq \frac{2(\mathbb{E}f(x_1) - f_*)}{\eta T} + \frac{1}{T}\sum_{t\leq T} \mathbb{E}[\|\xi_t\|^2 + s_t] \\ &\quad + \frac{\mathcal{O}_T^{\mathrm{J}} + \mathbb{E}B_T^{\mathrm{pot}}}{T} \\ &\leq \frac{2(\mathbb{E}f(x_1) - f_*)}{\eta T} + \frac{1}{T}\sum_{t\leq T} \mathbb{E}[\|\xi_t\|^2 + s_t] \\ &\quad + \frac{\mathcal{O}_T^{\mathrm{J}} + \mathcal{D}_T^{\mathrm{J}}}{T}. \end{aligned} \tag{37}$$

For $\kappa_t = \lambda c_t$, the cost on the left is $\lambda\mathbb{E}\sum_t c_t A_t/T$. Under the envelope over all feedback rules, the explicit bound on $\mathcal{D}_T^{\mathrm{J}}$ in Theorem A.5 applies.

Proof idea. Conditional unbiasedness about G_t separates acquisition error from reference noise and mean mismatch. Smoothness then bounds one-step descent by these three terms. Telescope, add expected weighted feedback cost, and apply Corollary A.8. The complete argument follows below. □

Here ξ_t includes reference bias and drift of the evaluation point, with $\|\xi_t\| \le L\|y_t - x_t\| + \|b_t\|$. The rule-learning bound controls the selective estimation term. The reference and the choice of evaluation point determine the noise and mismatch that remain.

Proof of Corollary A.9. Condition on the pre-query history $\mathcal{H}_t$ and the common reference gradient G_t. The independent Bernoulli query and the SRAO estimator give

$$\mathbb{E}[g_t \mid \mathcal{H}_t, G_t] = G_t, \qquad \mathbb{E}[\|g_t - G_t\|^2 \mid \mathcal{H}_t, G_t] = h_{t,0}.$$

When a query is certain, $g_t = G_t$ and $h_{t,0} = 0$, so the same identities hold. Taking conditional expectations over G_t and using $\mu_t = \mathbb{E}_t G_t$ yields

$$\mathbb{E}_t g_t = \mu_t, \qquad \mathbb{E}_t\|g_t - \mu_t\|^2 = \mathbb{E}_t\|G_t - \mu_t\|^2 + \mathbb{E}_t h_{t,0} = s_t + \mathbb{E}_t h_{t,0}.$$

The cross term vanishes because $\mathbb{E}[g_t - G_t \mid \mathcal{H}_t, G_t] = 0$. Adding the squared mean mismatch $\|\mu_t - \nabla f(x_t)\|^2 = \|\xi_t\|^2$ proves (36).

For the stated anchored update, smoothness and $\eta \le 1/L$ imply

$$f(x_{t+1}) \le f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|^2 + \frac{\eta}{2}\|g_t - \nabla f(x_t)\|^2.$$

Take expectations, telescope over the deterministic horizon, and use $f(x_{T+1}) \ge f_*$. The independent uniform choice of τ gives

$$\mathbb{E}\|\nabla f(x_\tau)\|^2 \le \frac{2(\mathbb{E}f(x_1) - f_*)}{\eta T} + \frac{1}{T}\sum_{t \le T} \mathbb{E}[\|\xi_t\|^2 + s_t + \mathbb{E}_t h_{t,0}].$$

Predictability of κ_t and the query law imply $\mathbb{E}_t[\kappa_t A_t] = \kappa_t \bar{p}_t$. Add the expected weighted feedback cost and apply the existing expected comparison from Corollary A.8,

$$\mathbb{E}\sum_{t \le T}[\mathbb{E}_t h_{t,0} + \kappa_t \bar{p}_t] \le \mathcal{O}_T^{\mathrm{J}} + \mathbb{E}B_T^{\mathrm{pot}} \le \mathcal{O}_T^{\mathrm{J}} + \mathcal{D}_T^{\mathrm{J}}.$$

This proves (37). The displayed moment and objective assumptions justify the expectations and finite-horizon summation. Finally, smoothness and the definition of ξ_t give $\|\xi_t\| \le L\|y_t - x_t\| + \|b_t\|$. □

For a clipping radius $C > 0$ and $\delta \ge 0$, the radial map $T_\delta(v) = v \min\{1, C/(\|v\| + \delta)\}$, with $T_\delta(0) = 0$, is nonexpansive. Under the same controller and integrability assumptions,

$$\mathbb{E}\sum_t \big[\|T_\delta(g_t) - T_\delta(G_t)\|^2 + \kappa_t A_t\big] \le \mathcal{O}_T^{\mathrm{J}} + \mathbb{E}B_T^{\mathrm{pot}}.$$

The comparison targets the same clipped reference and retains the comparator defined by raw gradient-estimation error. It does not require unbiasedness after clipping. Corollary A.10 gives the proof, including $\delta = 10^{-6}$ in real arithmetic. These distinctions determine the practical scope: the stationarity result specializes the update as stated, while clipping and AdamW in distillation or retained optimizer displacements in the agent implementation require empirical evaluation of task performance.

Corollary A.10 (Gradient-estimation error after clipping a common reference)**.** *Fix $c > 0$ and $\delta \ge 0$, and define $T_\delta(x) = x \min\{1, c/(\|x\| + \delta)\}$, with $T_\delta(0) = 0$. Under the protocol and integrability conditions of Corollary A.8,*

$$\mathbb{E}\sum_{t \le T}\big[\|T_\delta(g_t) - T_\delta(G_t)\|^2 + \kappa_t A_t\big] \le \mathcal{O}_T^{\mathrm{J}} + \mathbb{E}B_T^{\mathrm{pot}} \le \mathcal{O}_T^{\mathrm{J}} + \mathcal{D}_T^{\mathrm{J}}. \tag{38}$$

The oracle $\mathcal{O}_T^{\mathrm{J}}$ remains the oracle over fixed feedback rules for raw gradient-estimation error*, rather than clipped error.*

Proof. Put $f(r) = \min\{r, cr/(r + \delta)\}$ for $r > 0$ and $f(0) = 0$. It is continuous, $0 \le f(r) \le r$, and 1-Lipschitz: its inactive derivative is one, while on the active branch $r + \delta \ge c$, $0 \le f'(r) = c\delta/(r + \delta)^2 \le c/(r + \delta) \le 1$. For unit vectors u, v and positive r, s,

$$\|f(r)u - f(s)v\|^2 = (f(r) - f(s))^2 + f(r)f(s)\|u - v\|^2 \le (r - s)^2 + rs\|u - v\|^2 = \|ru - sv\|^2.$$

The zero-vector case follows from $f(r) \le r$. Thus T_δ is nonexpansive. Conditional on $\mathcal{H}_t, G_t$, contraction and Lemma A.2 give

$$\mathbb{E}[\|T_\delta(g_t) - T_\delta(G_t)\|^2 \mid \mathcal{H}_t, G_t] \le \mathbb{E}[\|g_t - G_t\|^2 \mid \mathcal{H}_t, G_t] = h_{t,0}.$$

At $\bar{p}_t = 1$ both sides vanish. Integrate over the potential reference gradient and history, use $\mathbb{E}_t[\kappa_t A_t] = \kappa_t \bar{p}_t$, and apply (34). □

For $\delta = 0$, T_δ is Euclidean norm projection. For $\delta = 10^{-6}$ it is the real-arithmetic native norm-clipping map. Finite-precision reductions and tensor multiplications are not covered by this exact inequality. The result does not require $\mathbb{E}[T_\delta(g_t) \mid \mathcal{H}_t, G_t] = T_\delta(G_t)$, and it does not replace the oracle for raw gradient-estimation error by a smaller oracle for clipped error. The target is the clipped version of the same potential reference gradient on the learner's encountered history, not a separately run Full trajectory. A realized B_T^{pot} cannot replace its expectation here. This gradient-estimation-error bound does not provide a convergence guarantee for clipped AdamW or a guarantee of held-out accuracy.

Fix $G = c > 0$, zero gradient predictions, query probabilities $(1/2, 1)$, equal weights on feedback rules and zero weighted feedback cost. Draw I from those weights and $A \mid I = i \sim \mathrm{Bernoulli}(p_i)$. Put $U = Ac/p_I$ and $V = 4Ac/3$. Here $V = \mathbb{E}[U \mid A, G]$ is the observable marginal. Let $h_i^\delta = \mathbb{E}[(T_\delta(U) - T_\delta(G))^2 \mid I = i]$ and $J_\delta = \frac{1}{2}(h_1^\delta + h_2^\delta) - \mathbb{E}[(T_\delta(V) - T_\delta(G))^2]$. For projection T_0 onto $[-c, c]$, $T_0(U) = T_0(V) = Ac$, yet

$$(h_1^0, h_2^0) = (c^2/2, 0), \qquad \mathrm{Var}_{i\sim r}(h_i^0) = c^4/16, \qquad J_0 = 0.$$

Thus no finite multiple of this difference in mean squared errors bounds the variation across rules' gradient-estimation errors. The clipped estimators are biased, with mean $3c/4$, so this is not a variance-reduction identity for unbiased estimators. For $z = \delta/c > 0$, direct evaluation of the three nonzero clipped values gives

$$\frac{J_\delta}{c^2} = \frac{z^2(4 + 12z + 6z^2)}{4(1+z)^2(2+z)^2(4+3z)^2} > 0.$$

Thus $J_\delta/(c^2 z^2) \to 1/64$. Since $h_2^\delta = 0$ and $h_1^\delta/c^2 \to 1/2$, we also have $\mathrm{Var}_{i\sim r}(h_i^\delta)/c^4 \to 1/16$ as $z \downarrow 0$. Exact zero is not claimed for a positive denominator offset. This excludes direct substitution of clipped gradient-estimation errors and this mean squared error difference into the bound for raw gradient-estimation error. It leaves the nonexpansion bound with its original oracle intact and gives no controller-failure, AdamW or accuracy conclusion.

B EXPERIMENTAL DETAILS AND SUPPORTING EVIDENCE

Table 3: **The same guidance–feedback principle yields different training updates.** The two applications specify their guidance, feedback and correction operator, with full feedback in agentic question answering and selective reference acquisition in on-policy distillation.

Component	Agentic question answering	On-policy distillation
Guidance	Reflection and backtracking	Previously computed gradients
Feedback	Terminal task outcomes	Solution-conditioned teacher targets
Evaluation point	After auxiliary updates	Current student state
Update	Projected auxiliary displacement and outcome step	Corrected gradient, norm clipping and AdamW
Acquisition	Full feedback	Fixed or learned selective feedback

B.1 CONTROLLED ESTIMATION AND ALLOCATION

This study separates the benefit of combining feedback rules from the quality of their gradient predictions. A common set of generated trajectories fixes the reference target for both estimators, and exact enumeration removes query-mask sampling error. A comparison with zero guidance then tests whether the supplied distillation gradient improves prediction at that same model state.

We evaluate Qwen3-0.6B and Qwen3-1.7B (Yang et al., 2025) at seeds 42 and 43 using the first eight questions of the GSM8K training split. Each question starts from the same initialized LoRA state (Hu et al., 2022) within its model–seed configuration. The design contains eight distinct questions, 32 model–seed–question groups and 256 generated responses. Questions are shared across the four model–seed configurations. LoRA rank is 64 and scaling is 128, with zero dropout and adapters on the query, key, value, output, gate, up and down projections. The trainable dimensions are 40,370,176 and 69,730,304, respectively. Model and adapter computations use FP32, TF32 is disabled, and norm/dot-product reductions use FP64. Updates are explicit parameter-coordinate SGD without momentum. The anchor is restored between comparisons and questions.

We use the OPSD student and privileged distillation prompts and generalized-JSD implementation (Zhao et al., 2026). The student sees the question and requests a boxed answer. The frozen, LoRA-disabled distillation policy q receives the reference solution and its conditioning prompt. Student thinking is disabled and distillation-policy thinking enabled. Full-vocabulary generalized JSD uses mixing parameter 0.5 and entrywise clipping at 0.05 before summing over vocabulary. Generation uses group size eight, temperature one, top-$p = 0.95$, top-$k = 20$ and at most 512 new tokens. One of the 256 responses reaches that cap and is retained. Numeric boxed-answer comparison supplies the binary reward.

After generating a group, we freeze its tokens, rewards, normalized advantages and old log probabilities. Advantages subtract the group mean and divide by its population standard deviation, floored at 10^{-6}. The outcome surrogate is the average token-level clipped GRPO loss (Shao et al., 2024), with symmetric ratio interval $[0.8, 1.2]$ and no additional reference KL. Let h be the mean auxiliary JSD gradient at the anchor x. The proposal is $a = -0.01h/\max(\|h\|, 10^{-30})$, and $y = x + a$. For each of the eight trajectories, we compute its outcome gradient u_i at y. Define $\bar{u} = 8^{-1}\sum_i u_i$, $P = h$ and $R_i = u_i - P$. Conditional on the completed group, a uniform trajectory index gives the exact target mean $\mathbb{E}R_i = \bar{u} - P$.

One reference packet is an outcome-gradient evaluation on an *already generated and rewarded* trajectory. Matching packet cost therefore isolates estimation efficiency conditional on this group. Generating and rewarding the trajectories, computing distillation guidance and enumerating the full gradients are additional costs of the controlled study. The result concerns the estimator at fixed states, before training trajectories can diverge.

Set $v = 8^{-1}\sum_i \|R_i\|^2$ and $s = 8^{-1}\sum_i \|u_i - \bar{u}\|^2$. We enumerate both Bernoulli-mask outcomes and every trajectory for $p \in \{1/8, 1/4, 1/2, 1\}$, verifying $\mathbb{E}\|P + AR_i/p - \bar{u}\|^2 = s + (p^{-1} - 1)v$. For fixed weights $(0.1, 0.2, 0.3, 0.4)$ on these four probabilities, $\bar{p} = 49/80$ preserves the expected gradient-packet cost. The rates and weights are shared by all 32 groups and fixed before response

Table 4: **Combining feedback rules reduces estimation error by about 40% at matched reference cost.** Conditional MSE is averaged over eight questions per model–seed configuration. Both estimators have expected reference-gradient cost $49/80$ and use the same states, guidance and outcome gradients.

Qwen3	Seed	Combined MSE	Sampled-rule MSE	Reduction
0.6B	42	20.87	34.56	39.6%
0.6B	43	7.16	11.87	39.7%
1.7B	42	1.10	1.83	40.1%
1.7B	43	0.96	1.60	40.1%

generation. The two conditional MSEs in Table 4 are $s + (\bar{p}^{-1} - 1)v$ and $s + v(\sum_a r_a/p_a - 1)$. Their difference is the predicted Jensen gain. Uniform weights provide a sensitivity comparison, yielding reductions between 45.00% and 45.40% across all groups and between 44.98% and 45.18% on the mixed-reward groups. Thus the combination gain persists under a second fixed weighting of the same query rates.

Eleven groups have mixed binary rewards. The other 21 have equal rewards and hence zero outcome advantages. In those groups the target gradient is zero even though the proposal direction and its correction residual can be nonzero.

The roughly 40% matched-cost reduction is a benefit of combining the same rules. The prediction $P = h$ itself is less useful than zero in this study. At the same evaluation point y, target and propensity, the chosen-minus-zero correction MSE is $(p^{-1} - 1)(\|h\|^2 - 2\langle h, \bar{u}\rangle)$. In all 21 equal-reward groups, the zero-guidance MSE is exactly zero, whereas $P = h$ adds query noise for $p < 1$. For the 11 mixed-reward groups, the prediction–reference cross moment is reconstructed from the dot product of the proposal and $-\eta\bar{u}$. This reconstruction uses FP32 vector operations and indicates negative prediction alignment at the evaluation point in every mixed group. At $p = 49/80$, the all-group MSE increases relative to the reconstructed zero-guidance value are 0.64%, 1.48%, 1.88% and 2.40% in the four table rows, respectively. Combining rules improves their estimation efficiency even with unfavorable guidance. Improving the guidance itself remains a separate requirement for further gains. The scalar study tests whether the estimation gain survives the cost of learning guidance and acquisition weights. It exposes two benefits: combining rules lowers the conditional error–cost objective on all three sequences, and the observed potential provides a much tighter finite-horizon certificate than the uniform maximum-based allowance. It uses four fixed feedback rules $(P_a, p_a) \in \{0, 1\} \times \{\varepsilon, 1\}$, $T = 1024$, and $\kappa_t = 1/2$. The potential reference gradient is $G_t = 0$ through round 256 and $G_t = 1$ afterwards. The third case replaces G_{513} by 32. The three constructions share one query seed and the same controller, which observes only queried reference gradients. Full knowledge of the synthetic sequence permits an offline comparison between the observed certificate and the underlying conditional error–cost objective.

Table 1 reports the conditional objectives. Table 5 explains their relationship to the learning guarantee. Here $\widehat{R}_T = \sum_t \widehat{f}_t - \min_a \widehat{L}_{T,a}$ compares observed inverse-probability-weighted losses. B_{obs} is the *additive* right side of (29), including the adjustment from known weighted feedback costs. $B_{\mathrm{pot}} = \Delta_T + \Delta_{T-1} - \sum_t \widehat{J}_t - \sum_t \zeta_t$ is the tighter observed certificate from the potential proof. B_{env} is the uniform *expected-remainder* bound (32). These quantities describe different comparisons. The observed certificate controls one weighted loss history, while the uniform bound controls an expected remainder.

Table 5: **The observed potential certifies small excess even when uniform allowances are loose.** All three observed excesses are covered by their potential certificates, which remain small relative to full-feedback loss. Units are squared residual plus weighted feedback cost, with $V = \max_{t,a} |G_t - P_a|^2$ and full-feedback loss $\kappa T = 512$. The two rate-one rules are duplicates in loss.

Construction	ε	V	$\widehat{R}_T$	B_{pot}	B_{obs}	B_{env}
Moderate floor	0.5	1	1.40	2.27	18.57	18.79
Small floor	1/8	1	−8.26	20.11	925.14	925.29
Single outlier	0.5	1024	0.56	2.40	9618.93	19238.03

Because Full is included, the observed certificate gives $\sum_t \widehat{f}_t \le \min_a \widehat{L}_{T,a} + B_{\mathrm{pot}} \le 512 + B_{\mathrm{pot}}$. The three displayed B_{pot} addends are approximately 0.44%, 3.93%, and 0.47% of this reference scale. The moderate case's best observed feedback-rule loss is 499.99, distinct from Full's 512. A small observed allowance can therefore certify a close comparison with the best fixed rule on that weighted history, while leaving the conditional estimation objective to be examined separately.

The conditional objectives in Table 1 show why the execution gain matters. With the smaller query floor, sampling the learned rules costs 541.00, above full feedback's 512. Conditional combination removes 46.55 of estimation error and brings SRAO to 494.45 without increasing expected query cost. The moderate construction also improves over full feedback. These comparisons attribute the gain to the executed estimator, with the learning history and weights held fixed.

The outlier separates this estimator advantage from the quality of the reference budget. SRAO still improves over random-rule execution, but its objective of 516.12 exceeds full feedback. A single residual of 32 also raises the uniform residual envelope to 1024. Smaller propensities and larger residuals magnify the maximum-based allowances, whereas the observed potential uses the losses encountered by the learner. Its small value certifies the observed weighted comparison, not the conditional objective or an expectation over query histories.

Adaptation is also gradual in the moderate case. The effective no-query gradient prediction first falls below 0.1 at round 4 and exceeds 0.9 at round 508, after the reference changes at round 257. The preserved no-query mass $\bar{q}_t = \sum_a r_{t,a}(1-p_a)$ determines how often the prediction affects execution, with no such contribution when $\bar{q}_t = 0$. These fixed-sequence examples explain the roles of query floors, residual scale and observed certificates. Evidence of fast tracking, switching regret or neural training gains would require different comparisons. The contextual study separates the availability of useful pre-query information from an online learner's ability to exploit it. The histories come from Qwen3-1.7B LoRA training on the GSM8K training split, with rank/scaling $64/128$, FP32 SGD at learning rate 10^{-3}, and a clipped-GRPO outcome objective. Each step computes solution-informed distillation guidance from one response and observes a fresh eight-response outcome group. The three histories held out from score fitting use auxiliary lookahead. The two outcome-only anchors evaluate guidance and feedback at the pre-update state and update only with the full reference gradient. Each history contains 100 steps.

Full-feedback histories make every reference available for comparisons on the same model trajectory. A two-feature hurdle score is fitted on full-feedback trajectories from seeds 2821–2823. With $\ell_t = \log L_t$ for pre-query guidance-completion length and $n_t = \log(\|h_t\|^2 + 10^{-12})$ for its guidance-gradient norm, the frozen rule uses the logistic sigmoid $\sigma(z) = (1 + \exp(-z))^{-1}$:

$$u_t \approx -20.78 + 4.00\ell_t + 0.85 n_t, \qquad m_t \approx 0.57 - 1.44\ell_t + 0.24 n_t,$$

$$s_t = \log \sigma(u_t) + m_t, \qquad p_t^{\mathrm{M50}} = \begin{cases} 1/2, & s_t \ge \tau, \\ 1/4, & s_t < \tau. \end{cases}$$

Here $\tau \approx -10.01$. Coefficients and threshold are shown to two decimal places. The score and threshold are fitted on the training trajectories and frozen before evaluation. Their inputs are pre-query guidance features.

Table 6: **Fixed contextual allocation improves the matched-cost variance comparison.** Positive d means lower gradient-estimation variance than a constant query rate at the same expected reference-token cost on that history. The improvement is normalized by total gradient energy. All five 100-step histories are included. The final row is exploratory.

Recorded history	Seed	d
Full-feedback lookahead, held out from fitting	2851	0.34
Full-feedback lookahead, held out from fitting	2852	0.05
Full-feedback lookahead, held out from fitting	2853	0.07
Outcome-only anchor, declared check	2860	0.19
Outcome-only anchor, exploratory check	2861	0.36

For each fully observed 100-step history, let c_t be its reference-batch token count and $v_t = \|G_t\|^2$. We compute

$$p_0 = \frac{\sum_t p_t^{\mathrm{M50}} c_t}{\sum_t c_t}, \qquad V_M = \sum_t ((p_t^{\mathrm{M50}})^{-1} - 1) v_t, \quad V_0 = (p_0^{-1} - 1) \sum_t v_t, \quad d = \frac{V_0 - V_M}{\sum_t v_t}.$$

The comparator p_0 matches expected reference-token cost in hindsight on the same history. The improvement d is normalized by gradient energy, rather than by V_0.

All five values of d are positive. The fixed rule assigns higher feedback probability to enough gradient energy to improve the matched-cost variance functional. The three held-out lookahead histories test this association beyond the trajectories used to fit the score. The outcome-only anchors show that it also appears when guidance does not displace the model before the reference is evaluated. Seed 2860 is the declared anchor check, and seed 2861 is exploratory. Guidance generation and gradient computation remain additional costs, so the variance comparison alone does not establish a net reduction in training work.

An online learner on seed 2860 produces a mixture with unit-packet surrogate $\sum_t [(p_t^{-1} - 1) v_t + p_t] = 67.24$, versus 62.59 for fixed M50 (7.43% higher), and token-matched $d = -0.11$. Mean execution still saves 3.54 relative to sampled-rule execution of those same weights. This history exposes three distinct effects. Context can support a useful fixed allocation rule, learning its mixture online can lose that advantage, and combining the resulting rules can still improve estimation at the learned weights.

Every reference group is observed, and the compared query probabilities leave the full-feedback updates unchanged. These conditional comparisons isolate allocation on a common trajectory. The trajectories induced by actual selective updates, together with the added cost of guidance, determine whether that allocation benefit translates into better end-to-end training.

B.2 On-policy distillation

The distillation study tests whether selectively acquiring reference gradients preserves the task-quality benefit of full-feedback training. Full feedback and SRAO with learned acquisition share the model, data, loss, optimizer and evaluation protocol. The fixed-step comparison uses training seeds 42 and 44 and evaluates all four step-100 models under one A800 inference profile. This design relates answer accuracy to the number of reference computations along each training trajectory.

We use Qwen3-1.7B and its tokenizer with the solution-conditioned distillation objective of OPSD (Zhao et al., 2026). The training pool contains 29,434 examples from the OpenThoughts mathematics split released with OPSD. Each run performs 100 optimizer steps with a logical batch size of 32, processing 3,200 training examples and generating 3,200 student responses over 100 logical steps. The student sees the problem. The teacher is the same base model with LoRA disabled and also receives the supplied solution. LoRA rank is 64 and scaling is 128 on the query, key, value, output, gate, up and down projections, with effective dropout zero. Student thinking is disabled and teacher thinking is enabled. Generation uses the OPSD templates, a 20,000-token prompt limit, a 1,024-token response limit, temperature 1.1, top-p 0.95 and top-k 20. The full-vocabulary forward-KL loss uses completion masks, temperature scaling and a 0.05 clamp on individual vocabulary contributions. This loss clamp is distinct from the parameter-gradient norm clip of 0.1.

Training uses BF16, FlashAttention2 and AdamW with learning rate 5×10^{-6}, betas $(0.9, 1 - 10^{-3})$, numerical constant 10^{-8} and zero weight decay. The linear scheduler has a 30-epoch horizon, with no learning-rate warmup and training stopped after 100 optimizer steps. Checkpoints are saved at steps 50 and 100. All methods train on the same RTX 4090 host using a logical batch of 32 examples, accumulated as eight groups of four. A reference gradient averages the eight group gradients, preserving each group's valid-completion-token denominator. One query acquires the reference for the entire logical batch. The single-GPU implementation and optimizer are shared by Full feedback and SRAO.

Learned acquisition applies Algorithm 1 to the cached raw reference gradient and zero, each paired with probabilities $1/8$, $1/4$, $1/2$ and 1. The cache changes only after a queried reference is computed, and stores the gradient before parameter-gradient clipping. The combined estimate enters the shared norm clip and AdamW update. At infinite learning rate, the recurrence retains equal weights over all minimizing rules. Full feedback computes the reference at every step. Fixed acquisition uses the same cache and correction with $p = 1/2$ and all weight on the cached-gradient rule. It omits the weight update in Algorithm 1. The zero-guidance setting uses the same rate library with zero predictions. These controls distinguish the complete learned-acquisition system from a fixed query rule and from acquisition without cached guidance.

Four initial full-feedback steps set κ to the mean squared raw reference-gradient norm and initialize the cache. Calibration leaves the rule weights unchanged, and its responses and reference computations enter the reported totals. Training seeds 42 and 44 use independent query-mask seeds 0 and 2. One uniform draw is consumed per logical step, including calibration, with a query when $u_t \geq 1 - \bar{p}_t$ after calibration. The cache and cost weight use training information only. Paired seeds share the training setup, while responses and model states can diverge as the methods update. The reference targets the distillation loss, so held-out answer accuracy provides a separate test of its usefulness.

Each step-100 model is evaluated on all 30 questions in AIME 2024, AIME 2025 and HMMT February 2025, with 12 sampled answers per question and vLLM model seed 0. We use the OPSD mathematics parser for answer extraction and correctness. Evaluation enables thinking, with temperature 1, top-p 0.95, top-k disabled, zero minimum probability and zero presence penalty. Context length is 40,960, with at most 38,912 new tokens per completion. The A800 profile uses tensor parallelism 1, memory utilization 0.9, CUDA graphs, at most 256 sequences and 8,192 batched tokens.

All 4,320 answers, including those reaching the length limit, contribute to the fixed-step comparison. We average correctness within each question, equally weight the three tasks within a seed, and then equally weight the two seeds. Appendix B.2 reports both seed-level and aggregate results. The two-seed point estimates describe the observed quality–reference trade-off. Accuracy equivalence or noninferiority would require a separate statistical design.

A reference batch counts one complete logical reference-gradient evaluation, including teacher inference and the reference backward computation. It is the unit controlled by the query decision. Student completion-token counts sum returned token-ID sequence lengths before padding or truncation, including any returned EOS or stop IDs and excluding prompts. These two measures distinguish selective reference computation from student generation. Teacher forward-token and effective-loss-token totals are unavailable, so reference-batch reductions measure gradient acquisition rather than teacher-token savings. Training time reflects the shared-host implementation as well as computation, while the calibrated squared-gradient weight κ controls the error–cost objective rather than pricing runtime.

Accuracy gained per acquired reference. Selective acquisition is useful only if the resulting model retains the benefit of training. Table 7 compares both trained systems with the same starting model under the A800 evaluation profile. Full feedback improves mean accuracy from 35.56% to 40.46%, and SRAO reaches 40.56% with 35.5% fewer reference-gradient batches. The corresponding gains of 4.91 and 5.00 percentage points show that reduced acquisition retains the observed benefit of distillation. Initial is evaluated once on all 90 questions with 12 answers each. Each trained model is evaluated on the same questions and sampling profile, and no checkpoint is selected within the step-100 comparison.

Table 7: **SRAO uses 35.5% fewer reference batches with similar observed mean accuracy.** Step-100 Avg@12 percentages are averaged over seeds 42 and 44. Initial is a single evaluation of the common starting model. SRAO uses learned acquisition. Mean equally weights the three tasks. Reference batches are summed over the two completed training runs, including four calibration batches per SRAO run.

Method	AIME24	AIME25	HMMT25	Mean	Ref. batches
Initial (Yang et al., 2025)	49.17	35.00	22.50	35.56	0
Full feedback (Zhao et al., 2026)	53.61	40.83	26.94	40.46	200
SRAO	54.31	39.03	28.33	40.56	129

Table 8: **Reduced reference acquisition preserves the observed mean accuracy across three training seeds.** Seeds 42 and 44 form the primary comparison. The seed-43 replication is shown separately. SRAO uses learned acquisition. Scores are Avg@12 percentages, the mean correctness of 12 answers per question. Mean equally weights the three benchmarks. Reference batches count complete reference-gradient evaluations during each completed 100-step training run, including SRAO's four calibration batches.

Seed	Method	AIME24	AIME25	HMMT25	Mean	Ref. batches
42	Full feedback (Zhao et al., 2026)	53.89	42.22	25.28	40.46	100
42	SRAO	53.61	38.89	29.17	40.56	65
44	Full feedback	53.33	39.44	28.61	40.46	100
44	SRAO	55.00	39.17	27.50	40.56	64
43	Full feedback	52.50	38.89	29.17	40.19	100
43	SRAO	55.56	41.11	26.94	41.20	64

The per-seed results in Table 8 show a small positive mean difference at both primary training seeds. The separate seed-43 replication increases mean accuracy by 1.02 points while reducing reference batches from 100 to 64. Task effects vary. For seeds 42 and 44, SRAO changes AIME24, AIME25 and HMMT25 by $+0.69$, -1.81 and $+1.39$ points, respectively. At seed 43, the corresponding changes are $+3.06$, $+2.22$ and -2.22 points. Thus the repeated mean trade-off does not imply uniform improvement across reasoning tasks. These fixed-seed point estimates support reduced reference acquisition with comparable observed means, without establishing accuracy equivalence.

Reference acquisition and response generation measure different work. SRAO reduces the first while preserving generation at every step. Table 9 shows that the two primary SRAO runs generate 5,658,746 student completion tokens, compared with 5,652,071 for full feedback, a 0.12% increase. Both produce 6,400 responses. Completion-token counts sum returned lengths before padding or truncation and exclude prompts. Teacher-token totals are unavailable, so the reference-batch reduction is not a measured token reduction. All accuracy comparisons use the same A800 inference profile.

Checkpoint selection tests whether the quality gain appears earlier. For each method, we screen steps 25, 50, 75 and 100 on the same ten questions per task with four answers per question. The score averages both seeds and all three tasks. One step is selected per method, shared across tasks and seeds, with ties resolved toward the earliest step. Full feedback has screening means 33.33%, 33.75%, 35.42% and 39.58%. SRAO has 31.67%, 37.08%, 37.08% and 37.08%. The rule therefore selects step 100 for full feedback and step 50 for SRAO. Fixed acquisition is not included in this checkpoint screen.

On the full benchmarks, the selected SRAO checkpoint improves the mean by 0.83 points with 61 rather than 200 reference batches (Table 10). Its differences are $+1.53$ points on AIME24, $+1.81$ on AIME25 and -0.83 on HMMT25. The mean is higher in both seeds, by 0.37 and 1.30 points. This comparison suggests that cached guidance and reference correction can preserve useful learning before the full-feedback budget is exhausted. Selection uses benchmark questions, and the complete full-feedback results were already available when the screen was designed. These are therefore benchmark-informed estimates rather than independent held-out validation. The full-feedback entries

Table 9: **Reference acquisition falls while student completion tokens remain nearly unchanged.** The comparison covers training seeds 42 and 44 on RTX 4090 GPUs. Reference batches and tokens cover completed 100-step trajectories, including calibration.

Method	Ref. batches	Completion tokens
Full feedback (Zhao et al., 2026)	200	5,652,071
SRAO	129	5,658,746

Table 10: **The selected SRAO checkpoint retains the accuracy gain with 69.5% fewer reference batches.** Full-benchmark Avg@12 percentages use one shared selected step per method and two training seeds. Reference counts include calibration and cover the training prefix ending at that step. The complete evaluation includes the questions used to select the checkpoint.

Seed	Method, step	AIME24	AIME25	HMMT25	Mean	Ref. batches
42	Full feedback (Zhao et al., 2026), 100	53.89	42.22	25.28	40.46	100
42	SRAO, 50	53.06	43.06	26.39	40.83	32
44	Full feedback, 100	53.33	39.44	28.61	40.46	100
44	SRAO, 50	57.22	42.22	25.83	41.76	29
Mean	Full feedback, 100	53.61	40.83	26.94	40.46	200
Mean	SRAO, 50	55.14	42.64	26.11	**41.30**	**61**

reuse the step-100 evaluations. All runs continued to 100 training steps, so the 69.5% saving describes checkpoint prefixes and excludes the work of checkpoint selection and evaluation.

Cached guidance and reference correction are effective with either acquisition setting. Fixed acquisition uses the latest reference as its prediction and queries with probability $1/2$ after four calibration steps. It is the single-rule specialization of equation (1). The zero-guidance setting instead learns the query rate with its predictions restricted to zero. Table 11 compares these settings at seed 44. Fixed acquisition attains 41.48% with 61 reference batches, compared with 40.56% and 64 for learned acquisition. Both retain the accuracy gain from training while reducing fresh reference computation.

Fixed acquisition is stronger on AIME24 and AIME25, while learned acquisition is stronger on HMMT25. Zero guidance achieves 40.09% with 60 references. The comparison supports the practical reference-correction construction but does not isolate a benefit from learning acquisition. Changing the guidance library or query rule also changes subsequent responses and model states. The mechanism studies complement these task results by holding states and expected reference cost fixed when assessing estimation quality.

Estimation and learning dynamics. The heterogeneous prediction library tests a benefit that shared predictions cannot provide: reducing disagreement between the estimates used without a query. We analyze three complete distillation histories at seeds 42, 43 and 44. Seeds 42 and 44 also supply the accuracy comparison in Appendix B.2. Each history contains four full-reference calibration steps and 96 adaptive rounds, giving 12 calibration and 288 adaptive rounds in total. This exploratory analysis connects the observed estimation geometry to the certificate scale on these training trajectories.

The library pairs $P_{t,a} = \alpha_a C_t$, for $\alpha_a \in \{0, 1\}$, with probabilities $\{1/8, 1/4, 1/2, 1\}$. Here C_t is the last acquired raw reference gradient. On each queried adaptive round, the moments $a_t = \|C_t\|^2$, $b_t = \|G_t\|^2$ and $d_t = \|G_t - C_t\|^2$ yield $\chi_t = (a_t + b_t - d_t)/2$ and

$$\|G_t - \alpha C_t\|^2 = b_t - 2\alpha\chi_t + \alpha^2 a_t.$$

These moments determine the rule errors, the two branches of J_t in (3), and the sharp bridge on every queried adaptive round. The calculation uses the raw, pre-clipping gradient geometry at the numerical precision of training.

On unqueried rounds, the reference and residual norms are unobserved. The observed quantities e_t, $\widehat{J}_t$ and $Q_t = \mathrm{Var}_{r_t}(e_t)$ are nevertheless zero by definition. Those rounds still enter the controller

Table 11: **Both SRAO settings use fewer references with higher observed means than full feedback at seed 44.** The comparison uses step-100 checkpoints and 30 questions per task, with 12 answers per question. Scores are Avg@12 percentages. Reference counts include four calibration queries per selective setting. These single-seed settings are reported separately from the two-seed means.

Method, seed 44	AIME24	AIME25	HMMT25	Mean	Ref. batches
Full feedback (Zhao et al., 2026)	53.33	39.44	28.61	40.46	100
SRAO, learned acquisition	55.00	39.17	27.50	40.56	64
SRAO, fixed acquisition	57.50	40.83	26.11	41.48	61
Zero guidance, learned acquisition	53.33	40.28	26.67	40.09	60

Table 12: **The no-query branch contributes about 15% of observed joint gain in each history.** The table decomposes gain and reports rule-loss variance over all 96 adaptive rounds per SRAO run. Both branch columns use the same importance weight as $\widehat{J}_t$. The 35, 36 and 36 unqueried rounds contribute zero observed feedback terms and remain included. Columns are rounded independently.

Seed	Queries	$\sum_t \widehat{J}_t$	No-query branch	Query branch	$\sum_t Q_t$
42	61	0.45	0.07	0.38	4.72×10^{-3}
43	60	0.66	0.09	0.56	0.58
44	60	0.58	0.08	0.50	0.28

recurrence through their known cost terms. The resulting branch decomposition tests whether combining distinct predictions contributes beyond combining query probabilities alone.

The no-query branch is positive on all 181 queried adaptive rounds and accounts for 15.58%, 14.35% and 14.42% of observed joint gain, respectively. This branch captures differences between the predictions, confirming a contribution that a common-prediction probability mixture would miss. The decomposition therefore connects the heterogeneous library to the conditional-weighting mechanism. Its estimation gain must still be distinguished from the quality of the learned allocation.

For that comparison, we evaluate the certificate hierarchy of Corollary A.8. The observed objective uses $\sum_t[\kappa\bar{p}_t + A_t h_{t,0}/\bar{p}_t]$ and the minimum accumulated observed error–cost over fixed rule identities on the same history. The best observed identity is $\alpha = 1, p = 1/2$ in each run. This comparator evaluates fixed-rule scores in hindsight on the SRAO trajectory. Training a fixed-probability policy would generally induce a different trajectory. Table 13 shows positive observed excess in all three histories.

The potential certificates cover those positive excesses, while the uniform allowance is much larger. For the zero/cache library, $M_t = \max(b_t, d_t)$, including rules with zero current weight. The acquired maxima of M_t/κ are 0.53, 4.56 and 3.96. Any uniform envelope V must cover these observations. At $K = 8$ and $\varepsilon = 1/8$, the coefficient multiplying V in (5) is approximately 1331.94. Thus the acquired residuals alone force the uniform allowance above the full-feedback cost scale 96κ by the factors in the final column. The unqueried residuals could enlarge the required envelope further. At this horizon, the uniform guarantee is too loose to certify an advantage over full feedback.

The four calibration queries per run initialize the cache and set κ from the mean squared reference norm. They enter the complete 100-step budgets and precede the adaptive certificate. Within each adaptive history, $\sum_t \bar{p}_t$ sums conditional query propensities along the realized path. Averaging acquisition across adaptive histories would require an expectation over their differing states and observations.

The two analyses assign distinct roles to the same feedback. Branch decomposition identifies a consistent estimation benefit from combining heterogeneous predictions. The certificates quantify the cost of learning their weights and reveal that the present uniform bound is uninformative at this horizon. The observed potential is more sensitive to the realized losses, but its single-history value cannot replace its expectation in the guarantee. The practical advantage in reference use is

Table 13: **Pathwise certificates cover the positive observed excesses, while the uniform allowance exceeds the full-feedback cost scale.** Each comparison uses one 96-round adaptive history. The final column gives a rounded lower bound on the uniform allowance for any residual envelope compatible with the acquired observations. It is not an upper bound on actual loss.

Seed	Observed excess	B_T^{pot}	B_T^{var}	B_T^{joint}	Allowance / 96κ
42	$+0.16$	0.43	2.64	14.42	7.41
43	$+4.43$	9.97	31.77	150.65	63.28
44	$+3.84$	8.52	26.14	125.52	54.97

Table 14: **SRAO reduces reference acquisition in the matched comparison, while published OPSD has the higher accuracy.** At 1.7B, the published OPSD mean is higher under a different protocol. All scores are Avg@12 percentages. Our lower blocks distinguish fixed-step and shared-checkpoint results, both averaged over seeds 42 and 44. Reference counts appear in Appendix B.2.

Model	Method	AIME24	AIME25	HMMT25	Mean
Published in OPSD, Table 2					
1.7B	Base (Instruct) (Yang et al., 2025)	51.5	36.7	23.1	37.1
1.7B	SFT (Zhao et al., 2026)	48.4	36.3	22.7	35.8
1.7B	GRPO (Shao et al., 2024)	51.1	38.3	23.7	37.7
1.7B	OPSD (Zhao et al., 2026)	57.2	43.9	29.2	43.4
Our fixed step-100 evaluation, seeds 42 and 44					
1.7B	Full feedback (Zhao et al., 2026)	53.61	40.83	26.94	40.46
1.7B	SRAO, learned acquisition	54.31	39.03	28.33	40.56
Our shared-checkpoint selection, seeds 42 and 44					
1.7B	SRAO, step 50	55.14	42.64	26.11	41.30

therefore tested by the task-quality comparisons, while the geometry explains which part of that design improves estimation.

Published comparison and checkpoint selection. Table 14 places the matched acquisition experiment in the context of published OPSD results (Zhao et al., 2026). The comparison is restricted to Qwen3-1.7B, the model used in our experiments. Published Avg@12 scores from OPSD Table 2 provide an external accuracy benchmark. Our matched full-feedback and SRAO runs test reference efficiency within the same objective, using either a fixed training step or one checkpoint selected jointly across tasks.

The two comparisons answer complementary questions. A fixed step holds the number of training updates constant, showing similar observed means with 35.5% fewer acquired references. Shared-checkpoint selection asks how much reference computation precedes a strong checkpoint. It selects SRAO at step 50 and full feedback at step 100, yielding 41.30% versus 40.46% with 61 versus 200 reference batches. This prefix comparison is benchmark-informed and does not include the work of completing or evaluating the training runs. Appendix B.2 gives the task effects and complete selection procedure.

Published OPSD's 43.40% remains 2.10 points above the shared-checkpoint SRAO result. Its reported scores are selected over checkpoints, whereas ours use one selected step shared across tasks and seeds. Training and inference implementations also differ. The published row therefore provides an external accuracy benchmark, while our matched full-feedback comparison quantifies reference efficiency. This distinction identifies both the demonstrated benefit and the remaining opportunity: preserving more of OPSD's peak task accuracy while reducing fresh reference computation. The published results have no measured reference-count coordinate for an accuracy–cost comparison.

B.3 AGENTIC QUESTION ANSWERING

The agentic study tests whether coordinating reflection guidance with terminal task feedback improves answer quality. This full-feedback implementation retains an auxiliary optimizer displacement after look-ahead and conflict projection. Outcome feedback is available at every training step, so the experiment examines the contribution of coordination within the complete training system.

Training. The agents are Qwen2.5-Instruct models at 3B and 7B (Qwen Team, 2025). Training uses HotpotQA (Yang et al., 2018) and 2WikiMultiHopQA (Ho et al., 2020), filtering trivial no-search and single-search questions. Qwen3-32B generates 600 successful SFT trajectories with reflection and at least five interactions, from questions on which Search-R1 failed. SFT construction can use reference answers. During RL, supplied answer aliases are redacted case-insensitively from the teacher tree. Detected aliases in reflection and backtracking spans suppress fast supervision. This masking does not cover every normalized alias or answer appearing elsewhere in a sampled prefix. The RL distillation policy is an exponential moving average of the optimized policy, distinct from the SFT generator.

The EMA policy q receives the active and archived execution tree, binary terminal success and structured hints. Fast supervision covers reflection and backtracking control tokens, including selected archived diagnostic branches. Its divergence is $\min\{\exp(-\delta) - 1 + \delta, 10\}$, with $\delta = \log \pi - \log q$. The slow objective combines group-normalized binary rewards, clipped GRPO (Shao et al., 2024) and KL regularization against π_{ref}. The regularization policy and distillation policy q have different roles. The slow loss covers generated tokens in the final active execution, excluding tool outputs and discarded branches. Trajectories are grouped by question before reward normalization.

Each outer step samples eight prompts and four trajectories per prompt. Three fast Adam steps form a candidate policy from the anchor. The same trajectories are rescored there before one slow Adam step. Conflict projection adjusts the fast displacement. The accepted update adds the slow displacement and weighted projected fast displacement to the anchor, retaining Adam moments across substeps. The EMA updates after the accepted policy update. Writing x_t for the anchor, y_t for the policy after the fast Adam steps and z_t for the policy after the slow Adam step, define $d_{f,t} = y_t - x_t$ and $d_{s,t} = z_t - y_t$. The update is

$$d_{f,t}^{\perp} = d_{f,t} - \frac{\min\{\langle d_{f,t}, d_{s,t}\rangle, 0\}}{\|d_{s,t}\|^2} d_{s,t}, \qquad x_{t+1} = x_t + d_{s,t} + w d_{f,t}^{\perp}.$$

Inner products sum stored trainable tensors across learner ranks. For $\|d_{s,t}\|^2 \le 10^{-12}$, the numerical guard leaves $d_{f,t}^{\perp} = d_{f,t}$. The selected configuration uses $w = 1$. Outside this guard, projection gives $\langle d_{f,t}^{\perp}, d_{s,t}\rangle = \max\{\langle d_{f,t}, d_{s,t}\rangle, 0\}$, so the retained fast displacement has a nonnegative inner product with the realized slow displacement in that geometry. This identity concerns the fused model tensors. The component study tests the complete coordination stage. The unbiased-gradient identity of selective acquisition applies to a different update construction. The EMA also receives terminal success when direct outcome optimization is removed.

Training uses 100 outer RL steps, LoRA rank 64 and scaling 128, Adam learning rate 10^{-6}, moments $(0.9, 0.98)$, clipping at one and weight decay 0.01. GRPO lower and upper clipping thresholds are 0.20 and 0.28, reference-policy KL weight is 10^{-3}, and context length is 16,384. Learner seed 1234 and data seed 42 specify one training configuration. The sensitivity sweep selects three fast Adam steps and projected-displacement weight $w = 1$ using the displayed seven-task mean F1. With $w = 1$, one, three and four fast steps give mean F1 scores of 41.26, 46.15 and 44.71. Because selection uses the reported QA tasks, these scores measure the selected fixed-seed configuration. The component comparisons include the training work associated with each configuration.

Evaluation and results. All retrieval methods share the November 1, 2023 English Wikipedia snapshot, retriever, top-three retrieval, context and tool-call budgets, decoding configuration and answer normalizer. Training reward is normalized exact match. Evaluation reports maximum-reference answer-token F1 in percent, including math transfer. The seven-task mean equally weights the two training domains and five additional tasks: Bamboogle (Press et al., 2023), FRAMES (Krishna et al., 2025), MuSiQue (Trivedi et al., 2022), Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017).

Table 15: **SRAO achieves the highest displayed F1 at both scales.** Answer-token F1 (%, ↑), with an equally weighted seven-task mean. One SRAO configuration per scale is selected by this mean. Retrieval methods share the retrieval and inference protocol. Base, FS-RAG and FL-RAG are implementation controls. Training data and computation differ across methods. Bold marks the highest F1 within each scale.

Method	2Wiki	Hotpot	Bamboogle	FRAMES	MuSiQue	NQ	TriviaQA	Avg.
Qwen2.5-3B-Instruct								
Base (Qwen Team, 2025)	23.98	24.08	9.45	8.01	9.70	14.27	38.84	18.33
CoT (Wei et al., 2022)	18.90	23.82	20.80	7.16	10.47	16.53	39.62	19.61
FS-RAG	15.47	25.85	10.48	10.42	7.64	19.84	45.38	19.30
FL-RAG	16.80	26.78	11.05	9.19	7.29	21.93	48.50	20.22
ReAct (Yao et al., 2023)	25.09	34.37	24.86	10.53	13.92	27.19	46.04	26.00
IRCoT (Trivedi et al., 2023)	15.89	24.50	25.27	6.79	12.43	27.86	49.19	23.13
TC-RAG (Jiang et al., 2025)	28.47	21.94	17.59	7.69	8.99	20.74	50.46	22.27
ReSearch (Chen et al., 2025)	27.23	33.96	15.09	10.00	9.47	34.61	53.93	26.33
Search-R1 (Jin et al., 2025)	29.90	37.24	29.90	10.76	13.53	34.73	55.08	30.16
AEPO (Dong et al., 2025a)	23.01	28.71	22.09	12.26	11.70	26.76	47.78	24.62
ARPO (Dong et al., 2025b)	29.55	36.48	27.32	13.49	13.38	33.29	53.66	29.60
MEM1 (Zhou et al., 2025)	18.06	20.15	5.19	4.99	4.47	19.18	33.09	15.02
SRAO	**48.01**	**56.17**	**35.25**	**23.51**	**31.02**	**55.37**	**73.72**	**46.15**
Qwen2.5-7B-Instruct								
Base	25.41	26.63	17.86	12.52	12.15	19.72	49.08	23.34
CoT (Wei et al., 2022)	23.55	29.10	37.56	17.60	14.35	22.47	49.33	27.71
FS-RAG	17.71	29.21	16.86	12.52	10.74	16.82	35.02	19.84
FL-RAG	19.78	34.42	24.10	12.10	12.46	19.72	42.66	23.61
ReAct (Yao et al., 2023)	27.51	42.81	27.63	15.29	19.34	30.01	54.55	31.02
IRCoT (Trivedi et al., 2023)	36.45	26.29	21.90	6.78	8.39	19.63	49.43	24.12
TC-RAG (Jiang et al., 2025)	29.70	40.83	25.13	16.46	17.56	29.01	54.78	30.50
ReSearch (Chen et al., 2025)	30.03	30.39	30.42	15.61	12.58	23.69	48.25	27.28
Search-R1 (Jin et al., 2025)	35.03	38.89	42.04	18.01	19.08	29.59	55.91	34.08
AEPO (Dong et al., 2025a)	19.88	13.85	13.24	7.24	5.85	9.93	17.53	12.50
ARPO (Dong et al., 2025b)	30.71	25.20	32.94	12.18	12.71	17.80	40.16	24.53
MEM1 (Zhou et al., 2025)	25.29	29.98	36.50	14.15	14.13	26.38	51.04	28.21
SRAO	**54.44**	**53.27**	**53.89**	**23.17**	**25.81**	**58.98**	**74.91**	**49.21**

Table 15 compares prompting, retrieval and reinforcement-learning baselines with implementation controls. Chain-of-thought prompting (Wei et al., 2022) supplies a reasoning baseline. ReAct (Yao et al., 2023) combines reasoning with tool use. The retrieval comparison includes TC-RAG (Jiang et al., 2025) and ReSearch (Chen et al., 2025). Search-R1 (Jin et al., 2025) provides an outcome-trained search agent. Other training baselines include AEPO (Dong et al., 2025a), ARPO (Dong et al., 2025b) and MEM1 (Zhou et al., 2025). FS-RAG and FL-RAG are fixed-sentence and fixed-length implementation controls. IRCoT (Trivedi et al., 2023) provides a sentence-level retrieval antecedent. Nearest-neighbor language modeling (Khandelwal et al., 2020) augments next-token probabilities with retrieved context–token neighbors. Differences in data, architecture and training work make these comparisons informative about complete systems.

Table 16: **The complete SRAO system has the highest seven-task mean, with task-specific trade-offs.** The 3B component study reports answer-token F1 (%, ↑). The distillation policy still receives binary outcomes without direct outcome optimization. Removing look-ahead coordination jointly removes rescoring, conflict projection and their associated training work.

Model	Method	2Wiki	Hotpot	Bamboogle	FRAMES	MuSiQue	NQ	TriviaQA	Avg.
3B	**SRAO**	48.01	56.17	35.25	23.51	31.02	55.37	73.72	**46.15**
3B	without reflection	26.41	37.74	18.35	11.46	18.18	45.28	58.45	30.84
3B	without backtracking	33.67	45.89	19.22	15.18	14.78	49.22	53.67	33.09
3B	without fast distillation	37.50	47.13	32.58	**25.00**	13.64	**63.33**	64.37	40.51
3B	without direct outcome optimization	30.12	55.29	25.29	21.43	20.73	54.65	66.28	39.11
3B	without look-ahead coordination	36.26	49.84	31.93	19.08	27.43	55.94	68.01	41.21

SRAO exceeds Search-R1's seven-task mean by 15.99 points at 3B and 15.13 points at 7B. Across the five tasks outside the named training domains, SRAO reaches mean F1 of 43.77 at 3B and 47.35 at 7B, compared with 28.80 and 32.93 for Search-R1. The corresponding differences are 14.97 and 14.43 points, computed before rounding. These results show that the full system's advantage extends beyond its two training domains, subject to the configuration-selection scope described above.

The component study in Table 16 identifies where that mean advantage is concentrated. Removing coordination lowers the mean by 4.94 points. Removing fast distillation or direct outcome optimization lowers it by 5.64 and 7.04 points, respectively. Coordination removal jointly changes rescoring, projection and their associated work, so the 4.94-point difference measures that bundled intervention. Task-level reversals qualify the aggregate gain. Removing coordination improves NQ from 55.37 to 55.94. Removing fast distillation improves FRAMES from 23.51 to 25.00 and Natural Questions from 55.37 to 63.33. Coordination and guidance thus help the seven-task mean while imposing trade-offs on individual tasks.

The evidence supports full-feedback coordination through the quality of the selected complete system and the mean effect of removing its component groups. The intervention measures the full coordination stage. Matched training budgets and independent seeds would be needed to separate its individual operations and assess variability. Selective-acquisition savings are evaluated by the distillation study, where the query decision directly controls reference computation.

C Related work

Knowledge distillation transfers information through a teacher's output distribution (Hinton et al., 2015). On-policy distillation applies that supervision to the student's own responses, addressing the difference between training responses and those produced by the student (Agarwal et al., 2024). OPSD uses solution-conditioned targets for on-policy self-distillation of reasoning models (Zhao et al., 2026). SRAO adopts this training objective in its distillation application. Its contribution there concerns the acquisition and correction of the reference gradient, while the teacher target continues to define the learning objective.

Guidance and outcomes also play complementary roles in reasoning agents. ReAct interleaves reasoning with actions (Yao et al., 2023), and IRCoT interleaves retrieval with intermediate reasoning for multi-step questions (Trivedi et al., 2023). Search-R1 trains reasoning and search behavior through outcome rewards (Jin et al., 2025). Process supervision supplies a different granularity of feedback by assessing intermediate reasoning steps (Lightman et al., 2024). SRAO's agent implementation combines reflection and backtracking guidance with terminal-outcome optimization. It uses the realized outcome displacement to determine which part of the auxiliary displacement enters the accepted update.

Several methods coordinate simultaneously available training signals. Gradient similarity can regulate auxiliary-loss contributions (Du et al., 2018). PCGrad projects away conflicting components between

task gradients (Yu et al., 2020). Conflict-averse gradient descent balances average descent with the worst local improvement across tasks (Liu et al., 2021), while GradNorm adjusts loss weights using gradient magnitudes and relative training rates (Chen et al., 2018). These methods motivate using update geometry to control guidance. SRAO's agent operator acts on optimizer displacements after auxiliary look-ahead and outcome rescoring. Its empirical evaluation is distinct from the selective reference-gradient guarantee, which requires predictions of a common reference.

Acquiring feedback selectively is also established in language-model training. Active preference learning chooses comparisons to improve the use of annotation budgets (Muldrew et al., 2024). Active-query RLHF studies how to obtain informative preference feedback during policy learning (Ji et al., 2025). More broadly, importance-weighted active learning corrects for selective label observation (Beygelzimer et al., 2009). SRAO considers a different acquired object: one complete reference gradient at the learner's current evaluation point. The residual of that observation both corrects the executed estimate and supplies feedback about every available prediction–probability rule.

The estimator combines familiar statistical ingredients. Inverse probability weighting compensates for unequal observation probabilities (Horvitz & Thompson, 1952). Augmented estimators add an auxiliary prediction and correct its residual when the target is observed (Robins et al., 1994). Multifidelity Monte Carlo uses correlated, lower-cost models to reduce the cost of estimating high-fidelity quantities (Peherstorfer et al., 2016). SRAO specializes prediction and residual correction to a shared training reference. Its conditional averaging is a Rao–Blackwellization of a randomly selected rule, conditioning on both the query outcome and any acquired reference. This classical projection supplies the estimation gain whose relation to learning variance is analyzed here.

Prediction-powered inference combines model predictions with observed labels to construct statistically valid estimates (Angelopoulos et al., 2023a). PPI++ tunes the prediction contribution to improve inferential efficiency (Angelopoulos et al., 2023b). In language-model evaluation, stratified prediction-powered inference allocates human evaluation across strata to improve the use of an annotation budget (Fisch et al., 2024). These works establish that the value of a prediction depends on how its error is corrected. SRAO carries this question into an evolving training process, where the target reference, available predictions and acquisition decisions can change with the learner's history.

Active statistical inference is especially close to the selective construction (Zrnic & Candes, 2024). Its augmented estimator has the same scalar residual-risk term $(p^{-1}-1)(Y-f(X))^2$, and its sequential formulation updates predictions and acquisition using collected labels. For mean estimation, oracle allocation is governed by the conditional root mean squared residual. Thus neither joint adaptation nor residual-based allocation is new here. SRAO's additional objects are a finite library of prediction–probability rules for a common training reference, their conditional aggregation across both query outcomes, and an expected cumulative comparison that includes reference acquisition price.

Prediction-powered methods also address optimization directly. Online power tuning learns a scalar prediction weight alongside model parameters using an unbiased gradient estimator (Shoham et al., 2025). Its labeled and unlabeled batch sizes are supplied at each iteration. PPI-SVRG uses a prediction-based snapshot control variate and relates prediction uncertainty to convex optimization error (Ao et al., 2026). SRAO instead learns weights over rules that specify both a prediction and a Bernoulli reference-query probability. Its main guarantee concerns executed estimation error plus acquisition price. The separate optimization consequence retains the supplied reference's bias and noise under its specified gradient update.

Adaptive importance sampling can reduce stochastic-gradient variance using feedback from optimization (Namkoong et al., 2017). Online variance reduction formulates the evolving sampling decision as an online learning problem (Borsos et al., 2018). VRM extends this approach to mixtures of sampling distributions and compares with the best fixed mixture in hindsight (Borsos et al., 2019). It already shares partial observations across mixture components along optimization-generated histories. SRAO acquires one complete reference through a Bernoulli decision and compares with a fixed rule identity selected outside the expectation. Different observation models and comparator classes make the corresponding horizon rates different statements. The sharp inequality here controls variation across joint prediction and acquisition rules through the error removed by their executed combination.

Aggregation and partial feedback have extensive online-learning foundations. Averaging expert predictions exploits the geometry of prediction losses (Kivinen & Warmuth, 1999), while adversarial bandit methods use importance weighting to learn from selected actions (Auer et al., 2002). SRAO's

observation model gives a queried reference to every rule, because all predictions refer to that same quantity and are available before acquisition. Learning additional rule weights therefore need not require additional reference queries. Constructing the predictions and computing their residual scores remain separate computational costs.

AdaHedge provides the adaptive mixability-gap learning rate and variance-sensitive penalty used by SRAO (de Rooij et al., 2014). Online learning with predictable sequences provides a broader basis for exploiting information available before an observation (Rakhlin & Sridharan, 2013). Second-order online bounds can also adapt between adversarial guarantees and stochastic fast rates (Koolen et al., 2016). SRAO incorporates the known current acquisition cost before choosing its weights. Its sharp inequality then connects the variance of the rules' error scores to the error removed by the conditional estimator. The required scale is the largest raw residual energy, which cannot uniformly be replaced by the largest rule risk.

Negative variance terms can compensate learning penalties and support selective observation (van der Hoeven et al., 2022). In that analysis, the negative term measures learner–comparator prediction disagreement under strong convexity and boundedness. This belongs to the wider study of conditions enabling fast statistical and online rates (van Erven et al., 2015). SRAO's negative term instead comes from conditioning out the latent rule across both query outcomes. The sharp relation makes this established compensation strategy apply to the stated reference-estimation and acquisition-cost problem. For a fixed library, positive query floor and bounded residuals, it yields a horizon-independent expected cumulative excess relative to a fixed rule on the learner's history. Sharpness concerns the one-round relation. The optimal query-floor dependence of the online comparison remains open.